Affective learning: improving engagement and enhancing learning with affect-aware feedback

Grawemeyer, B, Mavrikis, M, Holmes, W, Gutiérrez-Santos, S, Wiedmann, M & Rummel, N

Original citation & hyperlink:
Grawemeyer, B, Mavrikis, M, Holmes, W, Gutiérrez-Santos, S, Wiedmann, M & Rummel, N 2017, 'Affective learning: improving engagement and enhancing learning with affect-aware feedback', User Modeling and User-Adapted Interaction, vol. 27, no. 1, pp. 119-158
https://dx.doi.org/10.1007/s11257-017-9188-z

DOI 10.1007/s11257-017-9188-z
ISSN 0924-1868
ESSN 1573-1391

Publisher: Springer

The final publication is available at Springer via http://dx.doi.org/10.1007/s11257-017-9188-z

Noname manuscript No.
(will be inserted by the editor)

Affective learning: Improving engagement and enhancing learning with affect-aware feedback.

Beate Grawemeyer · Manolis Mavrikis · Wayne Holmes · Sergio Gutiérrez-Santos · Michael Wiedmann · Nikol Rummel

the date of receipt and acceptance should be inserted later

Abstract This paper describes the design and ecologically valid evaluation of a learner model that lies at the heart of an intelligent learning environment called *iTalk2Learn*. A core objective of the learner model is to adapt formative feedback based on students' affective states. Types of adaptation include *what* type of formative feedback should be provided and *how* it should be presented. Two Bayesian networks trained with data gathered in a series of Wizard-of-Oz studies are used for the adaptation process. This paper reports results from a quasi-experimental evaluation, in authentic classroom settings, which compared a version of *iTalk2Learn* that adapted feedback based on students' affective states as they were talking aloud with the system (the *affect* condition) with one that provided feedback based only on the students' performance (the *non-affect* condition). Our results suggest that affect-aware support contributes to reducing *boredom* and *off-task* behavior, and may have an effect on learning. We discuss the internal and ecological validity of the study, in light of pedagogical considerations that informed the design of the two conditions. Overall, the results of the study have implications both for the design of educational technology and for classroom approaches to teaching, because they highlight the important role that affect-aware modelling plays in the adaptive delivery of formative feedback to support learning.

Beate Grawemeyer & Sergio Gutiérrez-Santos
BBK Knowledge Lab, Birkbeck, University of London, London, UK
E-mail: {beate,sergut}@dcs.bbk.ac.uk

Manolis Mavrikis
UCL Knowledge Lab, UCL Institute of Education, University College London, UK
E-mail: {m.mavrikis}@ucl.ac.uk

Wayne Holmes
Institute of Educational Technology, The Open University, UK
E-mail: wayne.holmes}@open.ac.uk

Michael Wiedmann & Nikol Rummel
Institute of Educational Research, Ruhr-Universität Bochum, DE
E-mail: {michael.wiedmann,nikol.rummel}@rub.de

Keywords Affective learning · Bayesian networks · Formative feedback · Learner modelling

1 Introduction

The aim of our research is to enhance a student's learning experience and performance in a digital learning environment by providing intelligent support that goes beyond cognitive aspects and takes into account the student's affective state.

It is well understood that affect interacts with and influences the learning process (Kort et al., 2001; D'Mello et al., 2014; Baker et al., 2010). While positive affective states (such as surprise, satisfaction or curiosity) are known to contribute towards learning, negative affective states (including frustration and disillusionment) can undermine learning. For example, Woolf et al. (2009) describe how students can become overwhelmed (very confused or frustrated) during learning, which may increase their cognitive load (Sweller et al., 1998). In addition, Baker et al. (2010) found that certain types of affective states, such as boredom, were associated with poor learning and with gaming the system. However, when students are in a positive affective state, learning can be improved. For example, Csikszentmihalyi (1990) argues that students in a state of heightened engagement, that he calls *in flow*, are absorbed in the learning material and are thus primed for learning.

Any learning experience is typically full of transitions between positive and negative affective states. For example, while a student may be interested in a particular learning task, any misconceptions might lead to frustration or disillusionment as the student is forced to reconsider his or her existing understanding (in a process Piaget (1951) calls *accommodation*). However, if this negative affective state is reconciled, the student might once again become deeply engaged with the task. D'Mello et al. (2014), for example, elaborate how confusion, which initially might be thought of as a negative affective state, is likely under certain conditions to promote learning. It is important therefore, to deepen our understanding of the role of affective states for learning, and to be able to move students out of states that inhibit learning.

Pekrun (2006) discusses *achievement emotions*, affective states that arise in a learning situation and that are linked to learning, instruction, and achievement. In the *iTalk2Learn* project, we focussed on a subset of these achievement emotions: enjoyment (which we extend to *in flow*, by which we mean highly engaged, after Csikszentmihalyi, 1990), *surprise*, *frustration* and *boredom*. We also add *confusion*, which has been identified elsewhere as an important affective state during learning both for tutor support and for learning in general (Porayska-Pomsta et al., 2008; D'Mello et al., 2014).

Carenini et al. (2014) describe how effective support in learning situations needs to answer three main questions: (i) When should the support be provided? (ii) What should the support contain? And, (iii) how should the support be presented? In the *iTalk2Learn* project, the main support to be provided is

formative feedback, text messages that the system sends *during* the learning experience (Shute, 2008; Hattie and Timperley, 2007) in response to student problem-solving actions (Vanlehn, 2006) or other interactions (rather than messages that only summatively acknowledge the outcome of the learning experience).

Other research works, including our own (Mavrikis et al., 2008; Holmes et al., 2015a) and more recently that of Basu et al. (2017) in this issue, focus on how formative feedback in open-ended or exploratory environments can scaffold learners to perform a particular learning task. However, in addition to providing context-specific guidance, our formative feedback also aims to enhance student affective states - i.e., to move students from nominally negative affective states (such as frustration or boredom) into nominally positive affective states or to maintain positive affective states. In this context, we addressed Carenini's latter two questions by means of two Bayesian networks: one to determine *what* the feedback should contain (i.e. the *type* of formative feedback), the second for detecting *how* the formative feedback should be presented. Both networks were trained with data from a series of Wizard-of-Oz studies where we investigated the impact of feedback *type* and *presentation* on student affective states (c.f. Grawemeyer et al., 2015a,b). We learned that a student's affective state can be enhanced when the feedback *type* is matched to the affective state of the student. For example, when students were *confused*, affect boosts and specific instructive feedback were most effective. In addition, adapting the *presentation* of the feedback according to the students' affective state is also important, especially when the student is *confused* or *frustrated*. For these particular affective states, *high-interruptive* feedback (a pop-up window that has to be dismissed before the student can proceed) was more effective in enhancing the student's affective state, especially as the cost of not viewing the feedback is likely to be a negative affective state. However, when students were *in flow*, *low-interruptive* feedback (a glowing lightbulb which indicates that feedback is available) was preferred by students (Mavrikis et al., 2013).

While most research to date in this area responds to a student's affective state by adapting the feedback message (e.g. by including empathetic statements to motivate students (VanLehn et al., 2014; Forbes-Riley and Litman, 2011a; D'Mello et al., 2010), we instead adapt the *type* of feedback (whether it is, for example, instructional feedback or a reflective prompt), with the feedback content being based on the student's interaction (mainly their performance). In addition, instead of adapting the delivery of the feedback using for example an empathetic pedagogical agent (e.g. Conati and MacLaren, 2009; Rowe et al., 2009; Woolf et al., 2009), we adapt *how* interruptive the feedback is for the student (whether *low-* or *high-interruptive*).

In summary, in this paper we report on the development of intelligent formative support and its evaluation. The system includes a learner model that contains information about the student's affective state which is used to tailor the *type* of formative feedback and its *presentation* according to the student's affective state. It includes two Bayesian networks, one for each adaptation (feedback *type* and feedback *presentation*), which were trained with data from

earlier Wizard-of-Oz studies in which their predictive accuracy was tested. We subsequently built on that earlier work and incorporated the Bayesian networks in a comprehensive online system for the learning of fractions. In Rummel et al. (2016), we report on a summative evaluation of that system that focuses on learning efficacy. In this paper, we draw on elements of that summative evaluation to consider the potential of a *learner model* in relation to enhancing student affective states and learning. We evaluated the intelligent support and its learner model by comparing two conditions in real classroom settings. The first condition adapted feedback based on the student's affective state (the *affect* condition), whereas the second condition used only the student's performance (by which we mean the creation and manipulation of fractions representations) to provide feedback (the *non-affect* condition).

In the next section, we provide an overview of related literature. Section 3 describes the development of the affect-aware intelligent support. Section 4 outlines the evaluation of the support. Results of the evaluation are reported in Section 5. A detailed discussion that highlights the importance of affect-aware learner modelling is provided in Section 6, while Section 7 concludes the paper.

2 Related work

Different computational approaches have been adopted in order to detect affective states in intelligent learning environments. These include speech-based approaches (e.g. Cowie et al., 1999; Vogt and André, 2005), using information from facial expressions (e.g. Kaliouby and Robinson, 2004), keystrokes or mouse movements (e.g. Epp et al., 2011), or physiological sensors (e.g. Lang et al., 1993; Vyzas and Picard, 1998; Nasoz et al., 2003). Recent research (such as D'Mello and Graesser, 2010; Paleari et al., 2009; Wöllmer et al., 2010; Jiang et al., 2011) focuses on a combination of input stimuli to detect affective states.

Other research has investigated how a student's affective state or motivation can be detected or taken into account when providing appropriate learning material or motivational feedback. Early examples include del Soldato and du Boulay (1995) and Mavrikis et al. (2007) that look into a student's level of confidence and how much effort the student puts into performing a learning task as detected by the interaction with the learning environment (such as help requests or task completion).

Jaques et al. (2014) describe how they use gaze data to predict boredom and curiosity within *MetaTutor*, a hypermedia environment designed to foster student self-regulated learning processes in the domain of biology (Azevedo et al., 2009). Another example is Santos et al. (2014), which shows that personality and self-efficacy impact the effectiveness of motivational feedback and recommendations. Affective states were detected from mouse and keyboard interactions as well as from physiological parameters. Additionally, students self-reported their affective states through the Self-Assessment Manikin emotion assessment tool (Bradley and Lang, 1994) and free-text forms. The affec-

tive states that were detected included boredom, surprise, confusion, and loss of motivation. Wizard-of-Oz studies were used to investigate how motivational feedback and recommendations could be adapted based on students affective states.

Conati and MacLaren (2009) developed a model of emotions (a dynamic Bayesian network) based on students' bodily expressions in an educational game. The system used six emotional states: joy, distress, pride, shame, admiration and reproach. A pedagogical agent provided support according to students' emotional state detected by the system and their personal goal (such as wanting help, having fun, learning maths, or succeeding by oneself).

Another example is Shen et al. (2009), which also reports the use of Bayesian networks to classify students' affective states. Here biophysical signals, such as heart rate, skin conductance, blood pressure, and EEG brainwaves, are used for the classification. The detected affective states (interest, engagement, confusion, frustration, boredom, hopefulness, satisfaction, and disappointment) are included in an affective learner model. The system draws on the affective learner model and uses recommendation rules to determine appropriate interventions, such as providing an example when the student is confused or delivering a video/music when the student is bored.

Woolf et al. (2009) developed an affective pedagogical agent which is able to mirror a student's emotional state and alter the agent's feedback by providing for example, an empathetic message. These researchers used hardware sensors and facial movements to detect student emotions. This system discriminated between seven emotions: high/low pleasure, frustration, novelty, boredom, anxiety, and confidence. Different machine learning techniques were applied for the classification, including Bayesian networks and Hidden Markov models.

Similarly, Rowe et al. (2009) describe a narrative-centred learning environment, *Crystal Island*, which takes into account students' actions, locations, goals, and physiological information to detect their affective states. Naïve Bayes, decision trees, and support vector machines were used for the affect detection. The learning environment included virtual agents, which were able to express empathy based on the student's affective state. Another example is the *AutoTutor* tutoring system (D'Mello et al., 2005, 2010), which holds conversations with students in computer literacy and physics courses. The system classifies emotions based on natural language interaction, facial expressions, and gross body movements. The focus is on three emotions: frustration, confusion, and boredom. The classification is used to respond to students via a conversation through an embodied pedagogical agent and to adapt both the dialogue and the facial expression of the agent according to the student's affective state.

The promising results of the aforementioned work inspired us to investigate a related system that would fit our context. Methodologically, perhaps the most relevant work for our context is that of Forbes-Riley and Litman (2011a), who developed a physics text-based tutoring system, *UNC-ITSPOKE*. This used spoken dialogue (acoustic-prosodic and lexical features) to classify stu-

dent uncertainty. Based on the student's performance and level of uncertainty, the dialogue-based feedback is adapted. To develop their system, Forbes-Riley and Litman (2011a) used a corpus collected from a wizarded version of a spoken dialogue computer tutor (Forbes-Riley and Litman, 2011b), where detection and natural language understanding was performed by a human in order to train a model to detect uncertainty. Our aim was to extend this approach to accommodate several affective states (as previously investigated by D'Mello et al., 2010).

Meanwhile, several researchers have investigated adapting feedback to students' affective states. VanLehn et al. (2014) describe an affective meta tutor, which is able to determine what kind of motivational feedback should be provided to students based on their affective states and log data. Physiological sensors (facial expression camera and a posture-sensing chair) and a regression model are used to calculate whether a student is engaged, confused, or bored. A decision tree uses the current affective state of the student and log data to decide what motivational feedback message should be provided by an embodied pedagogical agent. In general, these adaptations tend to be focused on changing the text within a message (e.g. VanLehn et al., 2014) or changing the dialogue (e.g. Forbes-Riley and Litman, 2011a; D'Mello et al., 2010) to include, for example, empathetic statements designed to motivate the students. Interestingly, research has shown that there is a gender difference in how empathetic feedback is perceived by students - for example, Burleson and Picard (2007) show that female students respond more positively to empathetic feedback than male students; while Vail et al. (2015) showed that female students were more engaged and less frustrated when provided with affect-aware support than male students. Other research (e.g. Conati and MacLaren, 2009; D'Mello et al., 2005, 2010; VanLehn et al., 2014; Woolf et al., 2009; Rowe et al., 2009) have altered *how* feedback is delivered to students, for example through the use of a pedagogical agent capable of expressing empathy.

Finally, as explained in D'Mello and Kory (2015), most affect-aware systems up until now have been tested or evaluated only in lab-based contexts and very controlled settings.

In our research we extend the literature by (i) exploring how different *types* of feedback (e.g. reflective prompts or instructive feedback) can be adapted to a student's affective state, (ii) how the *presentation* of feedback can be adapted to a student's affective state by taking into account how interruptive that feedback is, and (iii) undertaking an ecologically-valid evaluation of an affect-aware learning environment in real classrooms.

3 The *iTalk2Learn* platform

Our research involves *iTalk2Learn*, an intelligent learning platform for children aged 8-12 years old who are learning fractions, which is designed to detect, analyse and respond to children's speech in real time in order to improve learning. Specifically, the platform's aim is to foster the robust learning of

fractions by providing activities that help develop conceptual knowledge, in an exploratory learning environment called *Fractions Lab*, which are interleaved with structured practice activities that help foster procedural knowledge, in an ITS called *Whizz Maths* (Mazziotti et al., 2015). The overall sequence of exploratory learning and structured practice activities is determined by a *Student Needs Analysis (SNA)* component. The SNA sequences the tasks according to the student's level of challenge, which is inferred from the student's interaction using the amount of feedback provided as a key indicator, in order to avoid students being over- or under-challenged, which may trigger boredom or anxiety (as described by Acee et al., 2010). In Rummel et al. (2016), we elaborate on the importance of interleaving exploratory learning and structured practice tasks and the potential of this interleaving for robust learning. In addition, in (Holmes et al., 2015a) we explore multiple dimensions of formative feedback provided while students are undertaking the learning activities in our exploratory learning environment. In this paper, however, we focus on how that intelligent support can usefully be made affect-aware (in Section 4.3 we explain how the sequence of exploratory learning and structured practice tasks was configured to address the aims of the evaluation described in this paper).

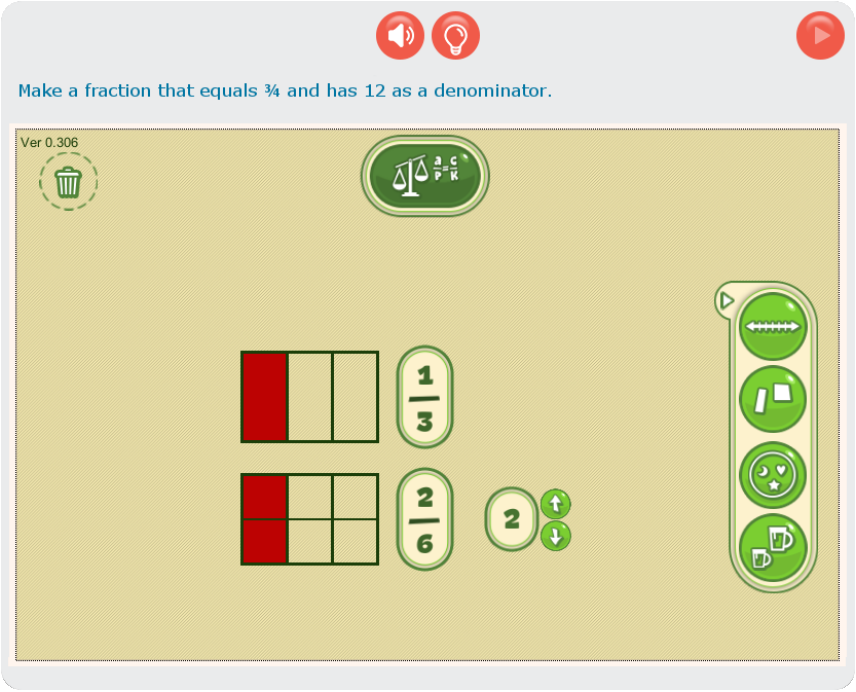

Fig. 1 Exploratory learning environment (*Fractions Lab*).

Figure 1 shows the *Fractions Lab* interface. Students are given a task (displayed at the top of the interface) which they explore and attempt to solve by choosing fraction representations (from the right hand side menu) which the student then manipulates (in the work area at the centre of the interface) in order to solve the given task. The large button at the top of the work area provides access to a variety of tools (to compare, add and subtract fractions). Adaptive feedback (which is well-known to be essential in exploratory learning environments, Kirschner et al., 2006) is provided to students based on their interactions with the system and their affective states.

Figure 2 shows the architecture of the adaptive support. Drawing on our previous work (Gutiérrez-Santos et al., 2012), the support comprises three main layers: the analysis layer, the reasoning layer, and the feedback generation layer.

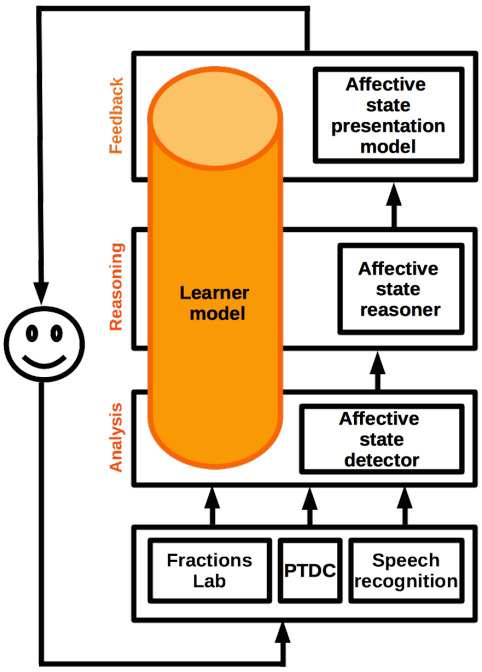

Fig. 2 Architecture of the adaptive support.

The **analysis layer** includes an affective state detector which has several inputs: a student's interaction with *Fractions Lab*, the output of a perceived task difficulty classifier (*PTDC*) which uses prosodic cues in the student's speech to predict the level of challenge for the current student, and the output from speech recognition software which identifies words in the student's speech. The analysis layer infers the student's affective state from these vari-

ous inputs. The student's detected affective state is then stored, together with the student's interaction data, in the learner model.

Building on the analysis layer, the **reasoning layer** decides *what* feedback should be provided. This layer contains an affective state reasoner, implemented as a Bayesian network which draws on information from the learner model, in particular the student's affective state, to decide *what type* of feedback should be provided to the student. The resulting feedback type is then stored in the learner model and provided to the feedback generation layer.

The **feedback generation layer** includes an affective state presentation model implemented as a second Bayesian network, which draws on the learner model to decide *how* the feedback should be *presented* to the student. The information that is accessed from the learner model includes the student's affective state as well as interaction data. The Bayesian network decides if the feedback should be provided in a *low-interruptive* or in a *high-interruptive* way.

We incorporated two Bayesian networks in order to accommodate the main architecture of the intelligent support.

When the feedback is provided depends on the student's interactions with the learning environment. If a student's inactivity passes a threshold time (which currently, based on the outputs of Wizard-of-Oz studies (Grawemeyer et al., 2015a,b), is set at 4 seconds), the intelligent support starts the reasoning process, drawing on the learner model, and calculates *what* type of feedback and *how* that feedback should be provided.

3.1 A hypothetical use case scenario

To illustrate the flow of information in the intelligent support, centred on the learner model, we next present a use case scenario.

Sarah, a primary school student who is learning about fractions, is using the *iTalk2Learn* system. Currently, she is working in *Fractions Lab*, exploring a task about the sum of two fractions, $\frac{1}{2}$ and $\frac{1}{3}$. She begins by creating a representation of $\frac{1}{2}$, an interaction which (like all interactions with representations and tools in *Fractions Lab*) is logged in her learner model. Sarah moves the mouse to create another representation, remarking as she does so "I think this's easy". The speech recognition component has also been continuously monitoring, transcribing what Sarah says and providing the words to the analysis layer. *Easy* is a keyword that is classified with a high probability of being *in flow*, and so the affect detection component classifies Sarah as *in flow* - more information that is saved in the learner model.

Sarah continues by creating a second fraction of $\frac{1}{3}$, but then stops, thinking about what to do next. This lack of interaction is noticed by the intelligent support which triggers it to start reasoning about her interactions and her affective state in order to deliver appropriate formative feedback. Based on the interaction data stored in the learner model (that reveals she has yet to receive any feedback) and her current affective state (*in flow*), the first Bayesian

network (the affective state reasoner) calculates that there is a probability of 1.0 that a *reflective* prompt will keep her *in flow*). Accordingly, the system chooses a *reflective prompt*, which asks her to reflect on her actions, "Why did you use this method?"

Next, again based on interaction and affect data stored in the learner model, the second Bayesian network (the affective state presentation model) calculates how the message should be presented. For Sarah, it determines that there is a 0.3 probability that providing feedback in a *high-interruptive* way will keep her *in flow*, whereas there is a probability of 0.7 that providing her feedback in a *low-interruptive* way (as an illuminated light bulb that she might or might not choose to access) will do so. Accordingly, the light bulb starts to glow (indicating that feedback is available). However, Sarah ignores the glowing light bulb - perhaps because she is *in flow*. Instead, she creates another fraction, $\frac{2}{5}$, suggesting that she has a misconception about how two fractions with different denominators are added together. Nevertheless, she puts all three representations into the *Fractions Lab* addition tool to check her calculation and then is confused to see that it is incorrect. She sighs, "This is so difficult…". Again, the speech recognition component provides the words to the analysis layer. *Difficult* is a keyword that is classified with a high probability as confusion, and so the affect detection component classifies Sarah as *confused* which again is saved in the learner model.

Meanwhile, Sarah continues to interact with *Fractions Lab*, exploring other ways to solve the task but, when she has run out of ideas, she stops again, unsure what to do next. Once more, this lack of interaction is noticed by the intelligent support and triggers it to start to reason about the information that had been stored in the learner model. The learner model reveals that she did not view and did not follow the most recent feedback and that her current affective state is *confusion*. Again, the first Bayesian network (the affective state reasoner) calculates which feedback *type* has the highest probability of improving her affective state. It determines that there is a probability of 0.6 that an *affect boost* will enhance her affective state, while there is a probability of 0.7 that *instructive feedback* will do so. Accordingly, the system chooses *instructive feedback* (instructing her to look at the denominators). Next, again based on interaction and affect data stored in the learner model, the second Bayesian network (the affective state presentation model) calculates how the message should be presented. It determines that there is a probability of 0.4 that providing feedback in a *low-interruptive* way will improve her affective state, whereas there is a 0.6 probability that providing feedback in a *high-interruptive* way will do so. Accordingly, this time, Sarah's instructive feedback message is presented in a high-interruptive pop-up window.

After other similar interactions, speech acts, Bayesian network calculations, and targeted formative feedback, Sarah finally completes her exploration of the task, and has discovered that $\frac{1}{2}$ plus $\frac{1}{3}$ equals $\frac{5}{6}$. At this point, she receives an affirmation prompt that acknowledges her success, and a final reflective prompt that takes into account the misconception detected by the system and asks her to reflect on what she has achieved. When she finishes this final reflection, a

Student Needs Analysis component (Mazziotti et al., 2015), determines that Sarah should move to the *Whizz Maths* environment, where she engages in a series of related structured tasks to practise what she had just explored and learned (the importance of matching denominators when adding fractions).

The following sections provide detailed information about the different components of the *iTalk2Learn* platform.

3.2 Learner model

The learner model spans all three main components and can be seen as the heart of the intelligent support. It includes the following information about the current student:

- Feedback data
  - The conditional probability table for the Bayesian network of the affective state reasoner, which is used to determine what *type* of feedback should be presented to students (please see Section 3.4 for more details).
  - The feedback messages that have been provided to the student.
  - The *type* of feedback provided to the student (e.g. *reflective prompts* or *instructive feedback*).
  - How feedback that was provided to the student was *presented* (*interruptive* or *non-interruptive*).
- Student data
  - The student's affective state (based on the student's speech and interaction and calculated by the affective state detector).
  - The student's progress with the task (whether the student is still exploring or has completed it).
  - The student's interactions with the learning environment (whether a representation has been created, selected or manipulated).
  - Whether or not the feedback was viewed by the student.
  - Whether or not the student followed the feedback.

The learner model is constantly being updated with information about the student and the feedback that has been provided to the student. The learner model is used by the various components to determine *what* type of support should be provided to the student and *how* that support should be provided.

3.3 Analysis layer (affective state detector)

The student's affective state is detected (inferred) from the student's speech and interaction with *Fractions Lab*. Data gathered in several Wizard-of-Oz studies (Mavrikis et al., 2014; Grawemeyer et al., 2015a,b) were used as the basis for our affect detection, as follows:

- The speech recognition software (Sail-Labs, 2016) detects whether students are speaking or not and produces an array of spoken words. This array is

used to detect keywords that are associated with a particular affective state. During the Wizard-of-Oz studies, we recorded what students said and used this to determine keywords that can provide some insight into the student's affective state. The selection of keywords was based on how often a particular word was spoken by the participating students when in a particular affective state and how unique the word was for that affective state. For example, from detected words such as 'that's easy', or 'this is good' the system infers the affective state of *in flow*, whereas from detected words such as 'this is hard', or 'tricky' the system infers *confusion*. The 'Bag of Words' method (e.g. Schuller et al., 2005) and a naive Bayes classifier was used to classify the student's affective state (Grawemeyer et al., 2014).
- What we call the perceived task difficulty classifier (*PTDC*), extracts prosodic features (such as 'um's and pauses) from the student's speech and uses speech and pause histograms to infer whether the student is under-, appropriately or over-challenged (Janning et al., 2014, 2016). The prosodic features were extracted from the voice recordings of the Wizard-of-Oz studies, based on two independent coders who classified a student's level of challenge by taking into account the student's speech and interaction with the learning environment.
- The student's interaction with the platform is used to add evidence towards an affective state. For example, whether or not the student viewed and followed the most recent feedback is used to calculate whether the student seems to be *in flow* or *confused*. For instance, if a student has viewed and followed the most recent feedback, the system infers that this student is *in flow*. However, if the student has viewed but not followed the most recent feedback, the system infers that the student is *confused*. In addition, the student's interaction is used in combination with the output of the PTDC to classify students as either *frustrated* or *bored*.

The affective state detector determines the student's overall affective state using weights given to the different inputs. Based on what was learned in our Wizard-of-Oz studies, the highest weight is given to the keyword detection, followed by PTDC and then interaction. Figure 3 shows a flow diagram of how the overall combined affect is calculated. For example, when a student has **not** viewed the most recent feedback, we infer the following affective states:

- *frustration*: (1) if the student is *over-challenged* and the interaction classification shows that the student is **not** confused (i.e. is *in flow*) and no keyword has been detected; or (2) if a keyword has been detected that is associated with *frustration*.
- *in flow*: (3) if the student is appropriately challenged and the interaction classification classifies the student as **not** confused and no keyword has been detected; or (4) if a keyword has been detected that is associated with being *in flow*; or (5) if the PTDC does not produce any results (when there was not enough speech data, or the speech data was too noisy, for the PTDC to infer the student's level of challenge), no keyword has been detected and the interaction classification classifies the student as *in flow*.

- *boredom*: (6) if the student is under-challenged and the interaction classification classifies the student as not confused and no keyword has been detected; or (7) if a keyword has been detected that is associated with *boredom*.
- *confusion*: (8) if the interaction classification identified the student as confused and no keyword was detected; or (9) if a keyword was detected that is associated with being *confused*.
- *surprise*: (10) only if a keyword was detected that is associated with *surprise*.

3.4 Reasoning layer (affective state reasoner)

The affective state reasoner uses the information from the student model to decide *what* type of feedback should be provided. We explore different types of feedback that are known from the literature (see Section 2) to support students in their learning and that fit our context: affect boosts, instructive feedback, other problem solving support, reflective prompts, talk aloud prompts, task sequence prompts, and affirmation prompts.

Table 1 shows an example feedback message for each feedback type.

The following different feedback types are provided while the student is exploring the task:

- AFFECT BOOSTS. As described in (Woolf et al., 2009), affect boosts can help to enhance a student's motivation to solve a particular learning task. Here, we included prompts that acknowledged that a task is challenging in order to encourage the student to keep trying. During the evaluation, affect boosts were provided only in the *affect* condition.
- INSTRUCTIVE feedback. This feedback provided detailed instructions about what action to perform in order to solve the task.
- OTHER PROBLEM SOLVING feedback. This aimed to help students tackle a problem by challenging their thinking instead of specifying next steps (a subset of 'Socratic' formative feedback, Holmes et al., 2015a).
- REFLECTIVE prompts. Reflecting on task performance and self-explanation can be viewed as a tool to help students address their own misunderstandings (Chi, 2000) and as a 'window' into their thinking.
- TALK ALOUD prompts. These build on the hypothesis that automatic speech recognition can facilitate learning, which is based mostly on educational research that has shown benefits of verbalization for learning (e.g. Askeland, 2012). During the evaluation, talk aloud prompts were provided only in the *affect* condition.

When a student has finished the task, the following additional feedback is provided:

- AFFIRMATION prompts. This feedback is provided when students have completed the task successfully, in order to indicate that they finished the task and should move to the next task.

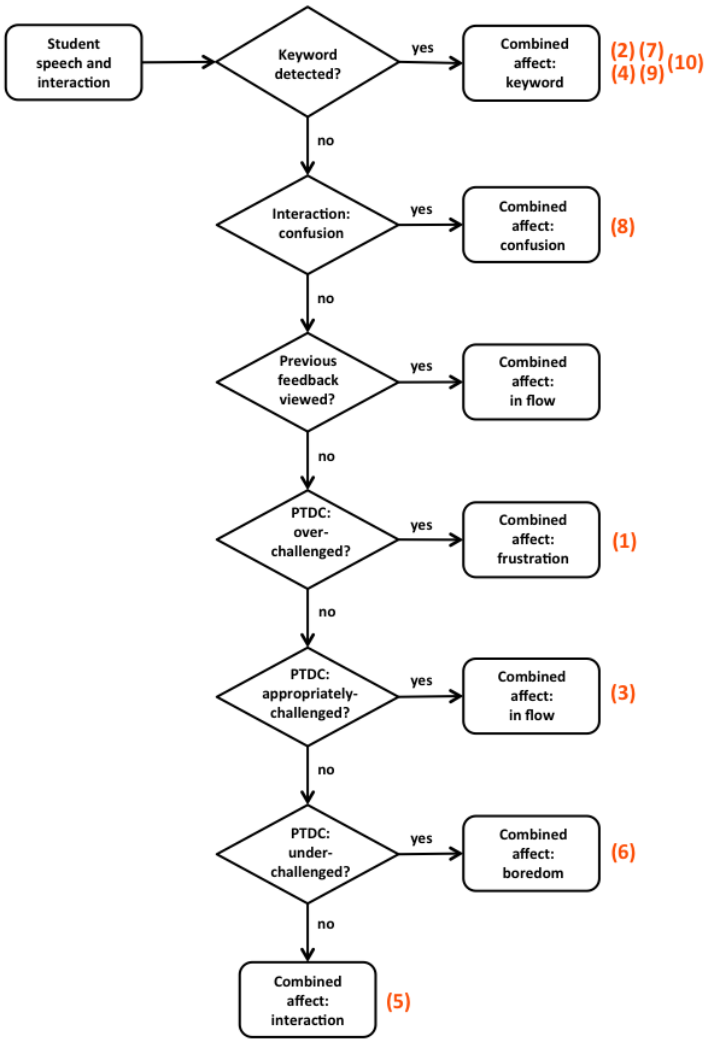

Fig. 3 Calculation of students affective state based on a combination of speech (keywords and prosodic features (PTDC)) and interaction.

– FINAL REFLECTIVE prompts that encourage students to reflect on a certain aspect of the task based on the students' performance.
– TASK SEQUENCE prompts. These are provided when students attempt to move to the next task without having completed the current task. Students are encouraged to first finish the current task or to ask for help but, in order to also allow them to proceed if they are stuck, students are able to move on to the next task with their third try.

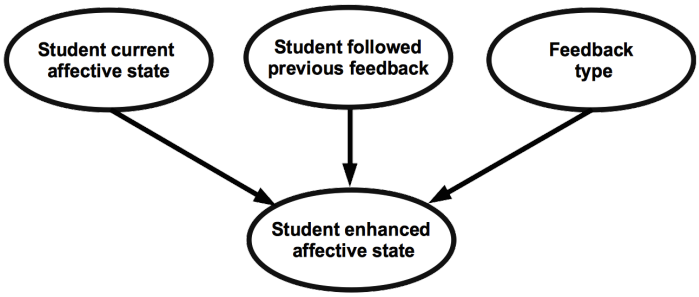

Fig. 4 Bayesian network of the affective state reasoner.

Based on the information from the learner model and the affective state reasoner, the system decides what *type* of feedback should be provided to the student.

Table 1 Examples of feedback types

Feedback type	Example
AFFECT BOOSTS	Well done. You're working really hard!
AFFIRMATION prompts	The way that you worked that out was excellent. Now go to the next task.
INSTRUCTIVE feedback	Use the comparison box to compare your fractions.
OTHER PROBLEM SOLVING feedback	What do you need to do now, to complete the fraction?
REFLECTIVE prompts	What do you notice about the two fractions?
TALK ALOUD prompts	Please explain what you are doing.
TASK SEQUENCE prompts	Are you sure that you have answered the task fully? Please read the task again.

The affective state reasoner is a Bayesian network based on data gathered in our Wizard-of-Oz studies (Grawemeyer et al., 2015b) that investigated the impact of the different feedback types on student affective states. In those studies, students were given a series of fractions tasks and were provided with feedback, of the types described above, by the researchers (the 'wizards') as if it was being provided by the system. The decision about what type of feedback to provide was based on a script. For more information, the reader is referred to Mavrikis et al. (2014).

Following these studies, we trained a Bayesian network using human-annotated data (265 data points) during and after the study as described in Grawemeyer et al. (2015b). Figure 4 shows the Bayesian network of the affective state reasoner on which we employed a 10-fold cross-validation that showed promising results (accuracy=79.25%; Kappa=0.50; recall true=0.62; recall false=0.87) and encouraged us to proceed to the full implementation of the system.

The affective state reasoner receives the current affective state of the student (based on the student's speech and interaction) as well as information about whether the student followed the most recent feedback. For each feedback *type*, the Bayesian network predicts whether the feedback type is able to improve a student's affective state. For example, enhancing a student's affective state from *frustration* to *confusion*, or from *confusion* to *in flow*. This is used to determine which feedback *type* will be most effective at enhancing the affective state.

Table 2 shows an extract of the conditional probability table (CPT) used by the affective state reasoner with example values. The CPT is stored in the learner model.

Table 2 Example extract of a CPT used by the affective state reasoner.

Student current affective state	Student followed previous feedback	Feedback type	Student enhanced affective state	
			FALSE	TRUE
in flow	F	affect boosts	0.3	0.7
confusion	F	affect boosts	0.4	0.6
frustration	F	affect boosts	0.3	0.7
boredom	F	affect boosts	0.5	0.5
surprise	F	affect boosts	0.5	0.5
in flow	F	instructive feedback	0.2	0.8
confusion	F	instructive feedback	0.3	0.7
frustration	F	instructive feedback	0.4	0.6
boredom	F	instructive feedback	0.3	0.7
surprise	F	instructive feedback	0.6	0.4
...				
...				
...				
in flow	T	reflective prompts	0.0	1.0
confusion	T	reflective prompts	0.6	0.4
frustration	T	reflective prompts	0.2	0.8
boredom	T	reflective prompts	0.5	0.5
surprise	T	reflective prompts	0.0	1.0

3.5 Feedback layer (affective state presentation model)

The aim of the affective state presentation model is to present the feedback in a way that enhances the student's affective state. In our learning environment, the feedback can be presented in either a *low-interruptive* way, by highlighting a light bulb at the top of the interface that indicates feedback is available that the student might or might not choose to access (see Figure 5), or in a *high-interruptive* way, by providing a pop-up window that has to be dismissed before the student can proceed (see Figure 6).

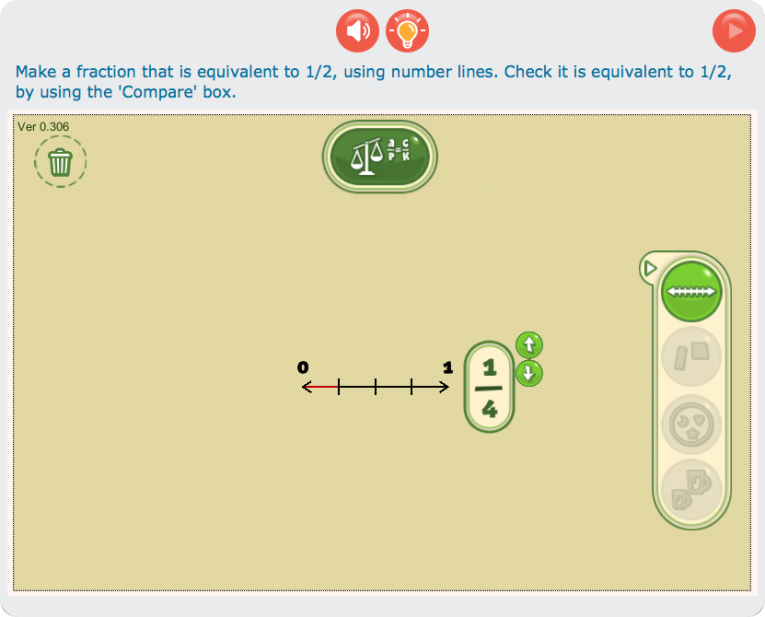

Fig. 5 Low-interruptive feedback. The light bulb at the top of the interface glowing in order to indicate that feedback is available.

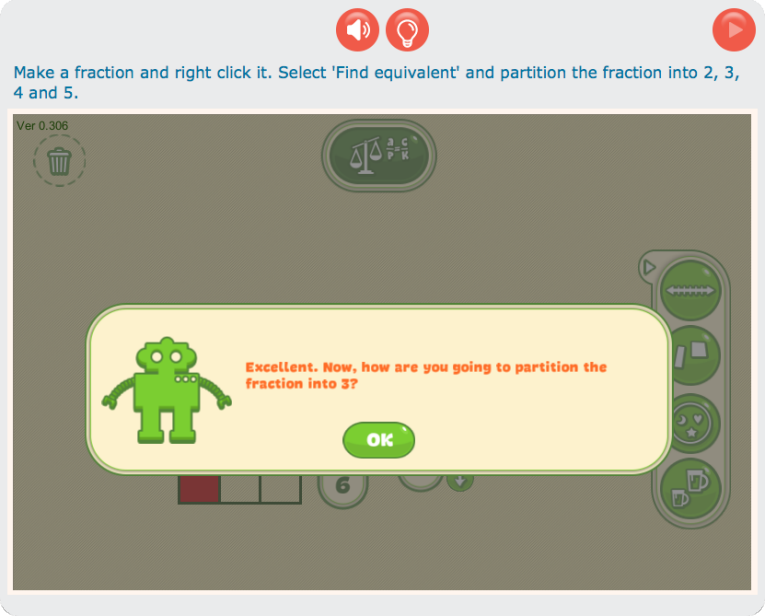

Fig. 6 High-interruptive feedback. A pop-up window that includes a feedback message.

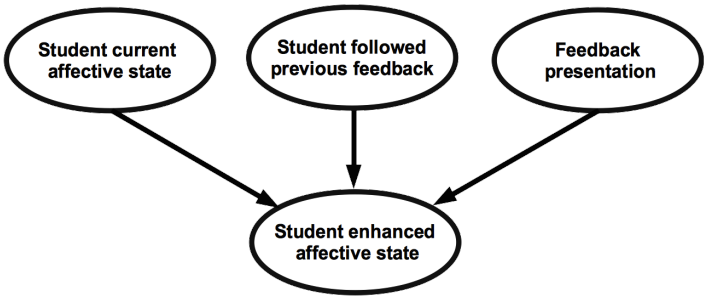

Fig. 7 Bayesian network of the affective state presentation model.

We conducted a further Wizard-of-Oz study that investigated if there was a difference in a student's affective state when the different types of feedback were either presented in the *low-interruptive* way, with the light bulb, or in the *high-interruptive* way, with the pop-up window (Grawemeyer et al., 2015a). Two independent researchers used video and speech recordings to annotate the students' affective states after the Wizard-of-Oz study (Kappa=.52, p<.001). The data from the study was used to train a Bayesian network that is able to predict whether the adaptation of the *presentation* of the feedback can improve a student's affective state. Figure 7 shows the Bayesian network of the affective state presentation model. This network is similar to the Bayesian network of the affective state reasoner (please see Figure 4), except that the node *feedback type* is replaced by the node *feedback presentation*. The dataset contained 266 cases. Each comprised the student's affective states that occurred before and after feedback was presented and the student's interaction data (whether or not the most recent feedback had been followed). As before, with this data set and employing a 10-fold cross-validation, we have encouraging results (accuracy=82.38%; Kappa=0.53; recall true=0.65; false=0.87).

The affective state presentation model receives the affective state of the student as well as information about whether the most recent feedback had been followed. Based on this, the *presentation* of the feedback most likely to enhance the affective state of the student is inferred.

4 Evaluation

As mentioned, we were particularly interested in the potential of our learner model for helping adapt support to promote student learning and engagement. The *iTalk2learn* project ran a series of formative and summative ecologically valid evaluations with students in real classrooms that considered a range of questions. In the summative evaluation, (among other questions that are tangential to this paper) we asked whether feedback that was tailored to a student's affective states enhanced the student's learning experiences and performance. To address this question, we evaluated the system and its intelligent support in a quasi-experimental study in which we compared one version that

included the affect-aware support (the *affect* condition) with a version where the affect-aware support was switched off (the *non-affect* condition).

We investigated the following sub-questions:

- Can a student's speech and interaction be used effectively as inputs in the detection of students' affective states?
- Is there a difference (between the *affect* and *non-affect* conditions) in how often feedback is accessed (based on the different feedback *presentation* mechanisms)?
- Are students in more positive affective states when feedback is tailored to their affective state?
- Are students less *off task* when feedback is tailored to their affective state?
- Do students have higher learning gains when feedback is adapted to their affective state?

4.1 Participants

80 students took part in the summative evaluation, although for several reasons our final dataset comprises 77 students. These participants were all primary school students, aged between 8 and 10 years old, recruited from two schools in the UK (one in the north of England, one in the south). Parental consent, for students' involvement in the study, was obtained for all participating students.

4.2 Conditions

The experimental condition (*affect* or *non-affect*) determined the presence or absence of the affective learner model. In turn, this determined what evidence was available to the analysis, reasoning and feedback layers (as described in Section 3), and the *type* of feedback and its *presentation*.

Table 3 summarises the differences in the *adaptation* mechanism and feedback *presentation* between the *affect* and the *non-affect* conditions. The difference between the conditions will be discussed below.

4.2.1 Affect condition

In the *affect* condition, the student's affective state was used to determine *what type* of feedback should be provided and *how* that feedback should be presented, in order to improve that affective state.

When students were working with *Fractions Lab*, the Bayesian model in the affective state reasoner (see Section 3.4) was used to provide the student with AFFECT BOOSTS, INSTRUCTIVE feedback, OTHER PROBLEM SOLVING support, or REFLECTIVE prompts. *How* this feedback was presented (*high-* or *low*-interruptive) was determined by the Bayesian network in the affective state presentation model (see Section 3.5).

Table 3 Feedback *type* and *presentation* in the affect and non-affect conditions

<table>
<tr><th></th><th colspan="2">ADAPTATION MECHANISM</th><th colspan="2">FEEDBACK PRESENTATION</th></tr>
<tr><td>FEEDBACK TYPE</td><td>AFFECT CONDITION</td><td>NON-AFFECT CONDITION</td><td>AFFECT CONDITION</td><td>NON-AFFECT CONDITION</td></tr>
<tr><td>Talk aloud</td><td>Delivered when the student has not spoken for 30 seconds.</td><td>n/a</td><td>High-interruptive</td><td>n/a</td></tr>
<tr><td>Affect boosts</td><td rowspan="4">Determined by the Bayesian model in the affective state reasoner.</td><td>n/a</td><td rowspan="4">Determined by the Bayesian network in the affective state presentation model.</td><td>n/a</td></tr>
<tr><td>Instructive</td><td rowspan="3">Rules take into account interaction and previous feedback</td><td rowspan="3">Low-interruptive</td></tr>
<tr><td>Other problem solving</td></tr>
<tr><td>Reflective</td></tr>
<tr><td>Reflective (final)</td><td colspan="2" rowspan="2">Delivered when the student has completed the task.</td><td colspan="2" rowspan="3">High-interruptive</td></tr>
<tr><td>Affirmation</td></tr>
<tr><td>Task sequence</td><td colspan="2">Delivered when the student presses the ‘next’ button but has not completed the task.</td></tr>
</table>

If students did not speak for 30 seconds, the system provided a TALK ALOUD prompt. This was always provided in a *high*-interruptive way as it was important in the *affect* condition that students talked aloud so that the system could detect the student's affective states from their speech.

When students attempted to move on to the next task (by clicking the ‘next’ button) without having finished the current task, the system provided a TASK SEQUENCE prompt (as described in Section 3.4). This prompt also was provided in a *high*-interruptive way to improve the chance that the student did not miss it (a *high*-interruptive prompt has to be responded to before a student can continue).

When students finished the *Fractions Lab* task they received an AFFIRMATION prompt and a final REFLECTIVE prompt. Both of those prompts were provided in a high-interruptive way in order to let the student know that the task is completed and to ask the student to reflect on their overall task performance.

4.2.2 Non-affect condition

In the *non-affect* condition, a narrower range of message types were provided than in the *affect* condition. In addition, *how* feedback was presented to students also differed.

Only student performance was used to determine the type of feedback provided to students in the *non-affect* condition, which meant that (because of the absence of the affective learner model) AFFECT BOOSTS were not provided and (because speech was not analysed) TALK ALOUD prompts were not provided. TALK ALOUD prompts were also not provided because asking

the students to talk without them perceiving any benefit could have been found intrusive by some students (Mavrikis et al., 2014; Grawemeyer et al., 2015a,b).

In the *non-affect* condition, INSTRUCTIVE feedback, OTHER PROBLEM SOLVING support, and REFLECTIVE prompts were provided in a *low*-interruptive way, as research (Mavrikis et al., 2013) has shown that students can find it very disruptive when they are interrupted during a learning activity.

TASK SEQUENCE prompts were provided when students attempted to move on to the next task but had not yet finished the current task (this is the same as in the *affect* condition). This prompt was presented in a *high*-interruptive way as a direct response to student's clicking the 'next' button as described above.

In the same way as in the *affect* condition, AFFIRMATION and final REFLECTIVE prompts were provided when students finished the task. These prompts were provided in a *high*-interruptive way in order to let the students know that the task is completed and to ask them to reflect on their overall task performance.

4.3 Procedure

The participating students were roughly stratified, according to previous teacher assessments of the children's mathematical ability, and then randomly allocated to two sub-groups (approximately equal in size, with each group having approximately the same number of high, middle and low achieving students). The first group (N=41) was assigned to the *affect* condition: the students were given access to the full iTalk2Learn system, which uses the student's affective state to determine the *type* of feedback and its *presentation*, as described above. The second group of students (N=36) was assigned to the *non-affect* condition: these students were given access to a version of the iTalk2Learn system in which feedback is based on the student's performance only.

Two series of sessions, one for each condition, were undertaken over several days in each school (in the school computer rooms, each of which was equipped with around 30 individual computers) at a variety of times of day which were balanced as far as possible between conditions. At the beginning of each session, students completed an online questionnaire that assessed the students knowledge of fractions (the pre-knowledge test - see Section 4.4 below). This was followed by 40 minutes during which the students engaged with fractions tasks.

For the purposes of this evaluation (to ensure that each student experienced a variety of exploratory and structured practice tasks), the SNA was configured to deliver two *Fractions Lab* exploratory tasks followed by four *Whizz Maths* structured practice tasks, a sequence that was repeated for the 40 minutes duration. The task provided to each student was based mainly on the student's performance in the previous task (which was calculated on the basis of the

amount of feedback provided: the more feedback provided in the previous task, the worse the student's performance was inferred to be). To enable students to proceed if they did not know what to do in a particular task, students were able to exit that task without finishing it but only after having been presented twice with a *TASK SEQUENCE* prompt that asked them to check and complete the task.

After the 40 minutes, students completed a second online questionnaire (the post-knowledge test) that again assessed their knowledge of fractions and also asked them about their experience using the system and emotional responses.

While students engaged with the system, a randomly allocated subset of students (affect condition: N=25; non-affect condition: N=22) sat at computers in the centre of the computer room in a way that allowed researchers to walk all around them, were monitored using the Baker-Rodrigo Ocumpaugh Monitoring Protocol (BROMP, Ocumpaugh et al., 2012). The researchers who undertook the coding, and who were trained in the BROMP method, recorded the student affective states and task behaviour data using the Human Affect Recording Tool (HART) Android mobile app.

The BROMP specifies strict guidelines for how affective states and task behaviour are detected. Each student is observed by a trained observer for up to 20 seconds. The student's body posture, facial expression and engagement with the learning environment are interpreted to infer whether the student is *in flow*, *confused*, *frustrated*, *bored*, *surprised*, or *delighted*. At the same time, the student's behaviour towards the task (whether the student is *off task*, or *on task*, or having an *on task conversation*, or having an *on task reflection*, or is *gaming the system*) is also monitored. At the end of the observation, the coder's interpretation of the student's affective state and behaviour is entered into the HART Android mobile app, and the researcher turns to the next student. The use of the app restricts the observers to the predetermined categories above (with the student being in *in flow* and *on task* as the default). However, when the affective state of the student was unclear to the observer, they are able to apply an unknown state (a questionmark in the app). This process is continuously repeated, thus logging multiple data points for each student, for the duration of the session.

4.4 Knowledge tests

Two isomorphic versions of six fractions problems were designed (see Figure 8). Students were randomly allocated one version at the first time of measurement (pre-test) and the other version at the second time of measurement (post-test). The students received one point for each correctly answered problem and consequently obtained an aggregated score that we used as an overall measure of fractions knowledge. Internal consistency of this scale was α=.57 at both pre- and post-test.

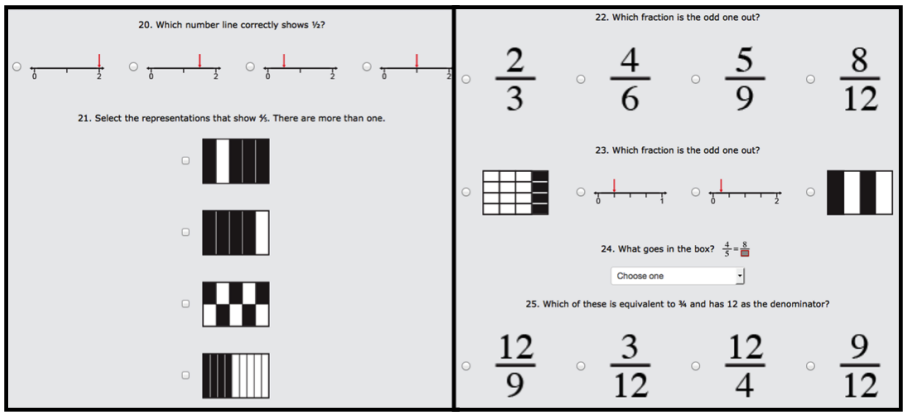

Fig. 8 Example extract from the pre- and post- questionnaires.

5 Results

5.1 Tasks provided to students

As described earlier, students were provided with a sequence of exploratory tasks (in the learning environment *Fractions Lab*) interleaved with structured practice tasks (in the learning environment *Whizz Maths*). In the *affect* condition, students engaged in 263 *Fractions Lab* tasks (M=6.41, SD=2.61) and 416 *Whizz Maths* tasks (M=10.15, SD=4.60). In the *non-affect* condition, students engaged in 293 *Fractions Lab* tasks (M=10.15, SD=2.92) and 450 *Whizz Maths* tasks (M=12.50, SD=4.81). Independent t-tests revealed significant differences between the conditions in the number of *Fractions Lab* tasks (t(75)=-2.738, p=.008, d=-.63) and the number of *Whizz Maths* tasks (t(75)=.-2.182, p=.032, d=-.50).

5.2 Affect detection

As described earlier, in the *affect* condition, a student's affective state was detected automatically by the system (henceforward, *automatic-detection*), by analysing the student's speech and interaction. In addition, the student's affective states were monitored and noted by two researchers using the BROMP method and HART mobile app (henceforward, *human-detection*). As described earlier, only a subset of students (*affect* condition: N=25; *non-affect* condition: N=22) were monitored in respect to their affective states.

The affective states that were both automatically and human-detected were *in flow*, *confusion*, *frustration*, *boredom*, and *surprise*. An additional affective state, *delight*, was human-detected. During the human-detection, the researchers were restricted to these 6 affective states (with the student being *in flow* as the default). However, when the affective state of the student was unclear to the researcher (during the human-detection), a "?" was annotated.

Both sets of data (from the automatic and human detection) include time stamps, identifying when a particular affective state was detected. This allowed the two sets of data to be matched (within a 10 seconds window). There was a moderate agreement between the automatically-detected affective states and the human-detected affective states, Kappa=.522, p<.001.

The difference between automatic- and human-detection was partly due to the affective state (*delight*) that was detected by the researchers and annotated with the HART tool but not detected automatically by the system. In addition, we knew from our Wizard-of-Oz studies that *surprise* and *boredom* are difficult to detect automatically. Excluding those affective states, there was a higher agreement between the automatically-detected affective states and the human-detected affective states, Kappa=.643, p<.001. However, this is lower than the commonly accepted Kappa threshold of .70 and it is important to note two caveats. First, if *delight*, *surprise* and *boredom* are excluded, we are ignoring some important aspects of human skills in affect detection. Second, the use of the BROMP protocol suggests that the annotated affective states may be less transient than they probably are. Nevertheless, given the authentic setting and our overall goal we consider this results acceptable but recognise that there is room for improvement. From a pedagogical point of view, we take into account that the effect of a misclassification will probably have a relatively low cost to a student's learning (first, a misclassification does not always lead to feedback being delivered or seen by a student and, second, any inappropriate feedback is unlikely to have a long term detrimental effect).

Table 4 displays the number of automatically-detected students affective states that matched the human-detected affective states and (in parentheses) the 'precision' of each detector i.e. the percentage of automatically-detected states that agreed with the human-detection. In our interpretation of the results we consider 'recall' (the percentage of correct detections of a state over the total number of cases in our dataset) even though this data is limited here because the 'gold standard' human-detected cases are sparse. Nevertheless, recall can help us judge the *relative contribution* of each component in the combined classification as well as the *relative completeness* of the whole system.

As mentioned earlier, keywords were used to detect all five different affective states whereas the PTDC component uses prosodic features to classify students as either under-, over- or appropriately challenged. The output from the PTDC were matched to the human-detected affective states as follows: under-challenged was matched to *boredom*, appropriately challenged to *in flow*, and over-challenged to *confused*. The interaction data included whether the student viewed the most recent feedback and whether the student followed that feedback. This information was used to calculate the probability that a student was either *in flow* or *confused* and, in combination with the PTDC, to determine whether students were *frustrated* or *bored* (as shown in Figure 3).

When keywords were detected, the match with the human-detected student affect was very high (*in flow*: 100.0% precision; *confused*: 96.7% precision).

Table 4 Comparison of human- and automatically-detected student affective states. The figure in parentheses gives the precision of the automatic detectors, the percentage of the automatically-detected states that were correct (i.e., matched the human-detected affective state).

	BROMP	Automatically-detected			
Affective state	annotation	Keywords	PTDC	Interaction	Combined
In flow	222	36 (100%)	139 (92.7%)	198 (82.5)%	187 (86.2%)
Confused	91	58 (96.7%)	87 (48.9%)	73 (59.8%)	79 (67.9%)
Frustrated	5	0 (0.0%)	n/a	n/a	4 (28.6%)
Bored	37	0 (0.0%)	8 (36.4%)	n/a	4 (80.0%)
Surprised	4	0 (0.0%)	n/a	n/a	0 (0.0%)

Unfortunately, however, as we further discuss in Section 6, keywords were not detected that often (recall *in flow*: 16.2%; recall *confused*: 63.7%). The detection of *in flow* by the PTDC (based on prosodic features of speech) was satisfactory (precision: 92.7%; recall: 62.6%). The precision of the detection of *in flow* from the interaction data was not as high (82.5%), however, the recall of *in flow* from the interaction data was high (89.2%). Combining the different sources did not lead to the highest precision or recall values for all of the different affective states, such as *in flow* (precision: 86.2%; recall: 84.2%). However, the combination of speech and interaction enabled the detection of *frustration* (precision: 28.6%; recall: 80.0%) and *boredom* (precision: 80.0%; recall: 10.8%).

5.3 Adapting the feedback *type*

In the two experimental conditions, different approaches were used to determine the *type* of feedback provided to the students. As described in Section 4.2, in the *affect* condition the feedback *type* was adapted based on the students' affective states as they answered the task, while in the *non-affect* condition the feedback *type* was based on students' performance. In addition, two feedback *types* (AFFECT BOOSTS and TALK ALOUD prompts) were only provided in the *affect* condition (please see Table 3).

In the *affect* condition, a total of 1971 feedback messages were provided to students (on average 48.07 messages per student, SD=14.58, min=25, max=92). In the *non-affect* condition, a total of 2007 messages were provided to students (on average 55.75 messages per student, SD=11.77, min=34, max=88). Figure 9 shows the different feedback *types* provided in each condition.

In order to investigate differences between the two conditions (*affect* and *non-affect*), a multivariate ANOVA was conducted for the different feedback *types*. Using Pillai's trace, there was a significant effect of the condition on the number of different *types* of feedback messages received, V=.929, F(5,71)=187.045, p=.000, η_p^2=.929. Separate t-tests on each feedback *type* were conducted. Alpha level was set to .01 following the Bonferroni correction for five comparisons. The t-tests revealed significant effects of adapting feedback *type* based on affect.

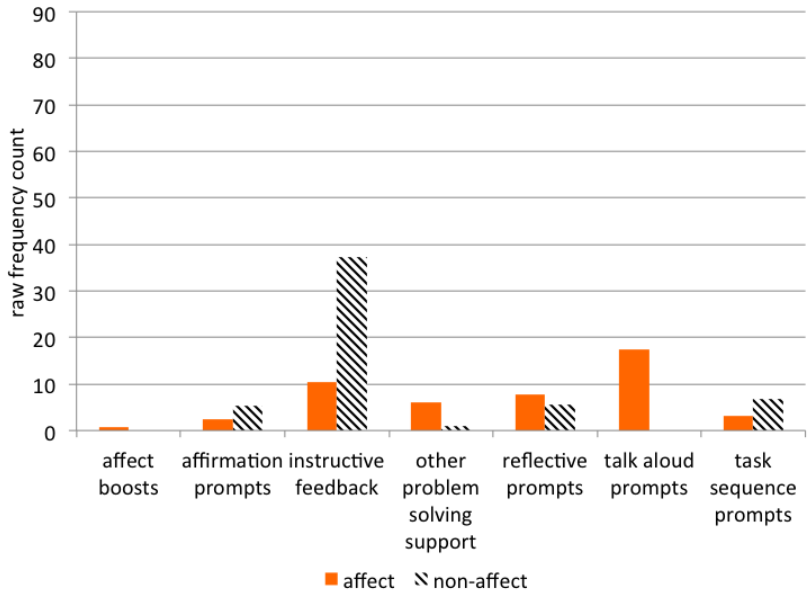

Fig. 9 Feedback *types* provided in the *affect* and *non-affect* condition.

AFFIRMATION prompts, INSTRUCTIVE feedback and TASK SEQUENCE prompts were provided less often in the *affect* condition than in the *non-affection* condition. In contrast, OTHER PROBLEM SOLVING support and REFLECTIVE prompts were provided more often in the *affect* condition than in the *non-affect* condition. As described earlier, AFFECT BOOSTS and TALK ALOUD prompts were only provided in the *affect* condition. See Table 5 for statistical details.

The reason why many more INSTRUCTIVE feedback messages were provided in the *non-affect* condition can be seen in the different aims of the two conditions. While the aim of the *non-affect* condition is to support students to reach a solution (based on the student's performance), the aim of the *affect* condition is not only to help student's solve the task, but also to improve a student's affective state. This can explain the larger amount of INSTRUCTIVE feedback messages in the *non-affect* condition, as the student is supported to reach a solution (or to perform a particular sub-task for this solution) through instructive feedback.

It is worth noting that intrinsic systematic differences between the two conditions, like differences in student affect, behaviour and/or learning, may have been introduced simply by monitoring and responding to affect rather than necessarily individualising the response to each student's affective states. We elaborate on this in Section 6.6.

Table 5 Statistical details of the feedback types provided.

Feedback type	Mean		Std. dev.		t-test
	affect	non affect	affect	non affect	
AFFECT BOOSTS	0.80	0.0	1.40	0.0	-
AFFIRMATION	2.51	5.33	2.09	2.41	t(75)=-5.50, p=.000, d=-1.25
INSTRUCTIVE	10.32	37.14	7.04	11.75	t(55.703)=-11.94, p=.000, d=-2.769
OTHER PROBLEM SOLVING	6.05	0.97	2.55	2.21	t(74.991)=9.36, p=.000, d=2.129
REFLECTIVE	7.80	5.53	3.49	2.21	t(68.501)=3.46, p=.001, d=0.846
TALK ALOUD	17.46	0.0	5.92	0.0	-
TASK SEQUENCE	3.12	6.78	2.60	4.22	t(56.679)=-4.50, p=.000, d=1.044

5.4 Responses to *low-interruptive* feedback

We were also interested in exploring whether there was a difference in the students' behaviour when offered *low-interruptive* feedback (i.e. whether or not the student clicked the light bulb). In the *affect* condition, students were provided with 389 *low-interruptive* messages (M=9.49, SD=3.551); while in the *non-affect* condition, students were provided with 1441 *low-interruptive* messages (M=40.03, SD=12.098) (the difference being due to the way in which the *low-interruptive* feedback was provided in the two conditions as explained in section 4.2). When feedback was *low-interruptive*, students could ignore the light bulb and therefore not see the feedback. In the *affect* condition, students ignored 74 of the *low-interruptive* feedback messages (M=1.80, SD=2.076). In the *non-affect* condition, students ignored 448 of the *low-interruptive* feedback messages (M=12.44, SD=10.814). In percentage terms, the students in the *affect* condition were more likely to view the low-interruptive feedback (81%) than students in the *non-affective* condition (69%). A t-test showed that this difference was a medium effect size and was statistically significant, t(75)=2.40, p=.019, d=.55. However, the difference in the rate of viewing and ignoring low-interruptive feedback between the two conditions may result from the greater frequency of low-interruptive feedback in the *non-affect* condition.

5.5 Affect and task behaviour

As described earlier, for a subset of students in both conditions (*affect* condition: N=25; *non-affect* condition: N=22) the students' affective states and task

behaviour were annotated by researchers using the Baker-Rodrigo Ocumpaugh Monitoring Protocol (BROMP) and the Human Affect Recording Tool (HART) Android mobile app (Ocumpaugh et al., 2012). This *human* detected affect data was used for further analysis as described below.

5.5.1 human-detected affect using BROMP

Figure 10 shows the different affective states that were annotated using the BROMP protocol.

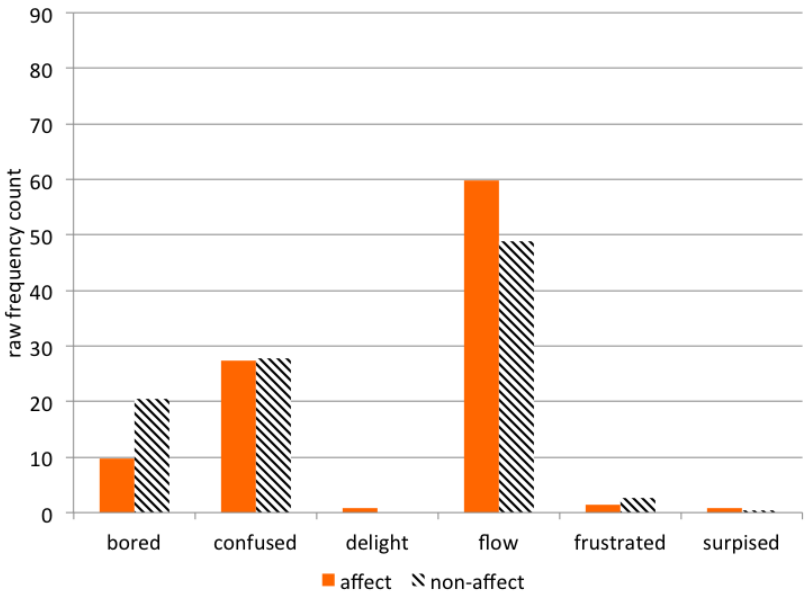

Fig. 10 Affective states annotated using the BROMP protocol (during the evaluation sessions in both conditions).

A multivariate ANOVA using Pillai's trace showed a significant **effect of adaptive support on the affective states** overall, V=.268, F(5,41)=3.006, p=.021, η_p^2=.268. In both conditions, students were mainly *in flow*. This was followed by *confusion* and *boredom*. Only rarely were students *frustrated*, *delighted*, or *surprised*. Follow-up t-tests showed that the effect of affect adaptation was statistically significant only for *boredom* (please see Table 6 for statistical details, statistically significant results highlighted in bold font), which was only half as frequent in the *affect* condition as in the *non-affect* condition.

5.5.2 Task behaviour using BROMP

Figure 11 shows the different task behaviours that were annotated using the BROMP protocol.

Table 6 Statistical details of students' affective states.

Affective state	Mean		Std. dev.		t-test
	affect	non affect	affect	non affect	
bored	**9.74**	**20.38**	**14.01**	**12.41**	**t(45)=-2.739, p=.009, d=-0.81**
confused	27.41	27.69	23.87	17.87	t(43.94)=-.046, p=.964, d=-0.01
delight	0.85	0.00	2.11	0.00	t(24.00)=2.011, p=.056, d=0.80
in flow	59.78	48.91	23.91	14.98	t(40.884)=1.891, p=.066, d=0.56
frustrated	1.48	2.67	2.81	4.66	t(45)=-1.075, p=.288, d=-0.32
surprised	0.74	0.35	1.52	1.64	t(45)=.843, p=.404, d=0.25

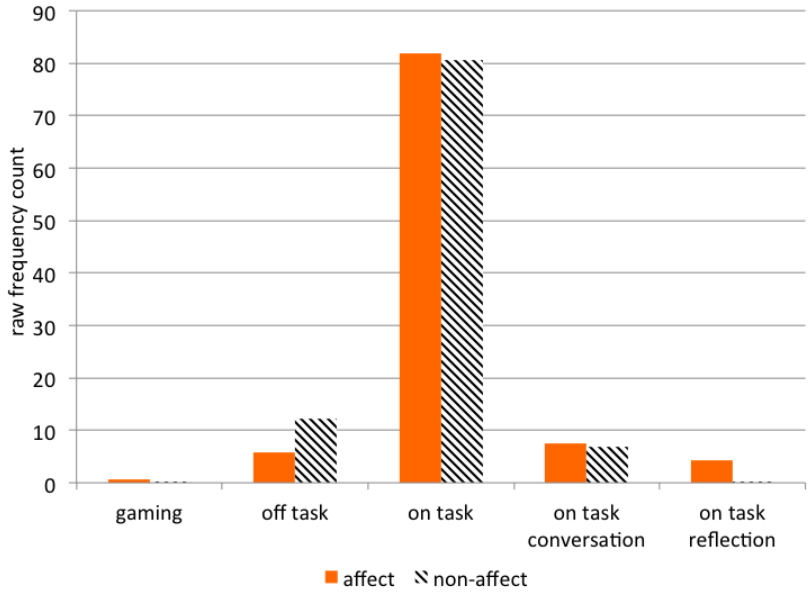

Fig. 11 Student task behaviour annotated using the BROMP protocol (during the evaluation sessions in both conditions).

A multivariate ANOVA using Pillai's trace showed a significant **effect of adaptive support on task behaviour** overall, V=.226, F(4,42)=3.071, p= .026, η_p^2=.226. In both conditions, students were mainly on task. Fewer students had an *on task conversation*, were *off task*, or *reflecting on the task*. Only seldomly were students *gaming* the system. Follow-up t-tests showed that the effect of affect adaptation was significant only for *off-task behaviour*.

Off task behaviour was only half as frequent in the *affect* condition as in the *non-affect* condition. Please see Table 7 for statistical details (statistically significant result highlighted in bold font).

Table 7 Statistical details of students' task behaviour.

Affective state	Mean		Std. dev.		t-test
	affect	non affect	affect	non affect	
gaming	0.67	0.12	2.31	0.55	t(24.00)=1.445, p=.161, d=0.58
off task	**5.75**	**12.31**	**6.95**	**8.75**	**t(45)=-2.862, p=.006, d=-0.84**
on task	81.87	80.67	14.92	10.65	t(43.268)=.320, p=.751., d=0.09
on task conversation	7.41	6.75	8.74	6.66	t(45)=.284, p=.778, d=0.08
on task reflection	4.31	0.27	12.29	1.25	t(24.568)=1.636, p=.115, d=0.60

5.6 Affect, task behaviour and performance

In the pre- and post-test questionnaire, students were scored according to how well they answered questions about fractions. In order to investigate if there was a relationship between affect, task behaviour and performance, we correlated the variables from the human-detected affect data (annotated with BROMP) with the post-test scores, while controlling for pre-test scores. However, there were no significant partial correlations of affect or task behaviour variables with the post-test scores.

5.7 Learning

Figure 12 shows the students' performance when answering fractions tasks before (in the pre-test) and after (in the post-test) they used the learning environment in the different conditions.

In the *affect* condition, students increased their knowledge of fractions from M=2.49 (SD=1.65) to M=3.83 (SD=1.46). In the *non-affect* condition, students increased their knowledge from M=2.44 (SD=1.58) to M=3.33 (SD=1.71). A repeated measures ANOVA showed a statistically significant increase of knowledge in both conditions (F(1,75)=43.94, p=.000, η_p^2=.369) but there were no significant differences between conditions at pre-test (t(75)=.12, p=.91, d=.03) or at post-test (t(75)=1.37, p=.17, d=.32), nor was there a significant interaction effect of time and condition (F(1,75)=1.81, p=.183, η_p^2=.024).

However, the observed tendency of the *affect* condition to show higher learning gains is promising and warrants further investigation.

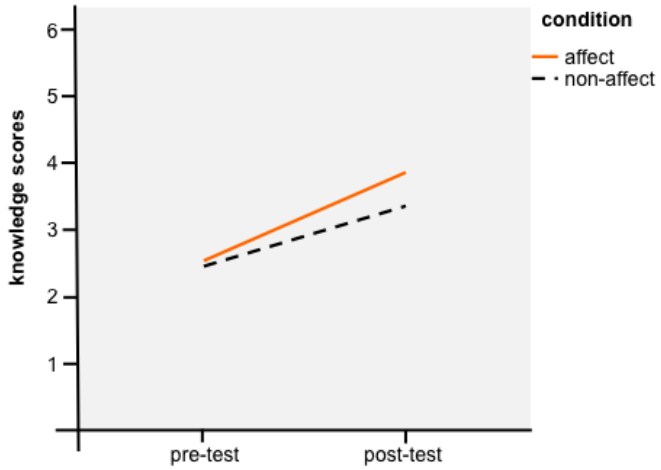

Fig. 12 Student learning gains in the *affect* and *non-affect* condition.

To explore a possible moderation effect of prior knowledge, we also calculated the conditional effects of condition on the post-test scores at the mean and plus/minus one SD from the mean of the pre-test scores (as suggested by Hayes, 2013). As can be seen in Figure 13, the effect of adapting support to affect is largest for students with low scores on the pre-test. However, because of the small sample size, we do not formally test this moderation model.

6 Discussion

The aim of our research is to enhance a student's learning experience and performance in a digital learning environment by providing intelligent formative feedback which takes into account students' affective states. This section discusses the results of our ecologically-valid evaluation in relation to our main research questions.

6.1 Can a student's speech and interaction be used effectively as inputs in the detection of students' affective states?

The automatic-detection of students' affective states was based on their speech and interaction with the learning environment. This was compared to the human-detected affective states that were annotated using BROMP with the HART mobile app (by observing the students' facial expressions, body posture and engagement with the learning environment). When taking into account all

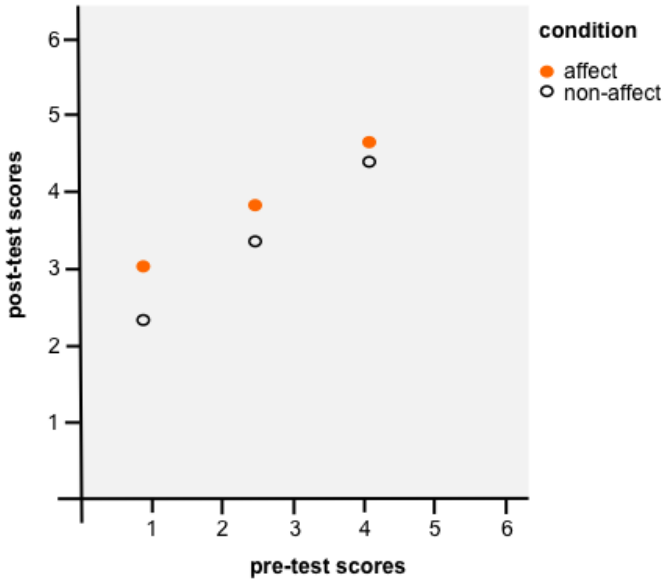

Fig. 13 Student post-test scores (y-axis) in the *affect* and *non-affect* condition for low, middle and high pre-test scores (x-axis).

of the automatically- and human-detected affective states (*in flow, confusion, frustration, boredom, surprise*, and the additional human-detected *delight*), the comparison revealed a medium agreement. The difference is mainly due to the quality of the automatic-detection, particularly in relation to *surprise* and *boredom*, which are difficult to detect automatically. This we knew from the Bayesian model's initial training, as there were far fewer instances.

However, when keywords were recognised in the students' speech, affect detection was very accurate (although unfortunately, in our ecologically valid setting, noisy real classrooms, keywords were recognised less often than in our previous lab tests). This is where speech recognition precision plays an important role and future research could aim to improve it.

As described earlier, the PTDC uses a student's speech to extract prosodic features (such as 'um's and pauses) to infer whether the student was under-, appropriately-, or over-challenged. As this classifier was not trained to detect affective states, it is not surprising that the precision of its detection of *confusion* or *boredom* was not high. However, although the recall of being *in flow* from the PTDC module was only 62.6%, this module does provide useful input for the overall detection of students being *in flow* (for which the overall precision was very high: 92.7%).

In contrast, detecting affective states from students interactions, revealed a high recall of being *in flow* (89.2%), but a lower precision (82.2%). The main advantage of detecting affective states from student interactions can be seen in combining it with the speech data to detect affective states that were difficult

to detect from speech only, such as *frustration* or *boredom*. For example, the precision of detecting *boredom* with the combined sources was 80.0%. However, the recall for *boredom* was only 10.8%. This is not surprising as we did learn from our earlier studies (as described in Section 5.2) that *boredom* is difficult to detect. Combining speech and interaction to detect *frustration* revealed a high recall (80.0%) but low precision (28.6%).

Nevertheless, overall, our analysis confirms the potential of the automatic-detection of affective states using information from speech (as identified in other work such as Forbes-Riley and Litman, 2011a) and that the combination of interaction data is promising and warrants further research.

6.2 Is there a difference (between the *affect* and *non-affect* conditions) in how often feedback is accessed (based on the different feedback *presentation* mechanisms)?

There was a difference between the conditions in how often *low-interruptive* feedback was accessed (by clicking the light-bulb). More feedback was ignored by students in the *non-affect* condition (a result that was statistically significant). The reason for this might be that students who were in a particular affective state, such as *confusion* or *frustration*, might have not realised that feedback was available. Grawemeyer et al. (2015a) describe that when students are in a particular affective state, such as *frustration*, *low-interruptive* feedback might be ignored because of cognitive load. In contrast, the *presentation* of the feedback in the *affect* condition took students' affective state into account.

However, the relatively large number of *low-interruptive* feedback messages in the *non-affect* condition compared to the *affect* condition (N=1441 vs N=389) might have had some additive effects such that, after a point, students became 'immune' and systematically disregarded the *low-interruptive* feedback messages.

6.3 Are students in more positive affective states when feedback is tailored to the their affective state?

In both conditions students were mainly in positive affective states rather than negative affective states. Similar to Conati and MacLaren (2009), the reason might be the nature of the learning environment (in our case, the exploratory nature of *Fractions Lab*). However, there was a statistically significant difference in how often students experienced *boredom*, with the students in the *affect* condition being bored less often than students in the *non-affect* condition. The reason for this can be found in the way the feedback was adapted. Students in the *affect* condition received feedback that varied more often than did those in the *non-affect* condition. In addition, the variation of the feedback *types* within the *affect* condition was directly in response to student affect. Finally, by adapting and altering how the feedback delivery interrupted the student

(*interruptive* or *non-interruptive*), students who are in negative affective states can be supported to move into a positive affective state. In contrast, when feedback is *not* adapted based on a student's affective state, there is a risk that feedback is ignored (especially and perhaps more critically when students are in a negative affective state).

While there are probably individual differences in the effectiveness of emotional support (Santos et al., 2014), our trained Bayesian networks were able to enhance a student's learning experience by reducing *boredom*. However, it would be interesting to explore whether there are individual differences in the effectiveness of our different feedback types on a student's affective states and learning.

6.4 Are students less *off task* when feedback is tailored to their affective state?

Students in both conditions were mainly *on task*, which again might be explained by the nature or novelty of the exploratory learning environment, which appeared to engage the students.

However, we found a difference between the conditions in *off task* behaviour. Students in the *affect* condition were less *off task* than students in the *non-affect* condition, a result that was statistically significant. Here, the adaptations of the feedback *type* as well as the adaptation of the feedback's *presentation* based on the student's affective state had an effect on their engagement with the task. As discussed in Baker (2007), off-task behaviour is likely to be related to affective states. Hence, adapting feedback to a student's affective states can improve the student's task behaviour.

In addition, results from Section 5.1 showed that students in the *affect* condition encountered fewer exploratory (*Fractions Lab*) tasks and fewer structured practice (*Whizz Maths*) tasks, results that were statistically significant. The results also show that students in the *affect* condition made fewer attempts to exit an exploratory task without finishing it (clicking the 'next' button and receiving a TASK SEQUENCE prompt, as described in Section 5.3), which again was statistically significant. This suggests that students in the *affect* condition were more engaged with the tasks and spent more time on solving the task instead of trying to move quickly on to the next task without finishing it. This also supports our result that students in the *affect* condition were less *bored* and also less *off task*.

6.5 Do students have higher learning gains when feedback is adapted to their affective state?

Student knowledge of fractions was improved in both conditions, but the difference between the conditions was not statistically significant. Woolf et al. (2009) show that a student's *on task* behaviour leads to higher post-test scores, which,

given that our results show that students in both conditions were mainly *on task*, might explain why students in both conditions improved their knowledge of fractions. Nevertheless, students in the *affect* condition had slightly higher learning gains, which although not statistically significant is encouraging.

Our results support Forbes-Riley and Litman (2011a) who describe how the affect-aware version of ITSPOKE led to higher learning than a control condition, although the difference was only statistically significant for a subset of students (those who received the most uncertainty adaptations).

Our results also showed that the difference between the conditions in respect to learning gains was highest when students had low pre-test knowledge scores. This supports research from D'Mello and Graesser (2013), which showed that learning gains improved for students with low baseline knowledge when the system responded to *confusion*.

As discussed above, the aim of our affect-aware support is to enhance a student's affective state - to move them from negative to positive affective states - in order to improve student learning. Interestingly, our findings suggest that students who had low pre-knowledge benefitted more from the affect-aware support. This is presumably because these low-attaining students experienced more negative affective states than those students who had high pre-knowledge, and thus there were more opportunities for the system to move them to a positive affective state. This warrants further investigation.

6.6 Limitations

The limitations of this study are due mainly to practical constraints and our decision to test the system in a setting as ecologically valid as possible.

The first two limitations are technical. First, we deliberately restricted ourselves to voice and limited interaction data only, our goal being to test how far these relatively straightforward and easily scalable modalities could be taken. Unlike physiological sensing (such as facial expression or galvanic skin responses monitoring), voice and interaction data are perceived by teachers and parents as less intrusive and require only basic technology in the classrooms: voice data only requires a microphone, which are common in many schools, while the interaction data requires no additional client-side technology. In fact, given the promising results from our research (we were able to detect effectively students' affective states automatically based on their voice and interaction data only), this limitation is also a strength.

Second, although we developed the exploratory learning environment (*Fractions Lab*) and therefore were able to control its intelligent support components, the structured practice environment (the pre-existing *Whizz Maths*) was effectively a black box. Presumably, more direct access to structured practice interaction data would have enabled a richer and more accurate affect diagnosis. Nevertheless, there is merit to black-boxing the learning activity and establishing a protocol of communication between that part of the system and the affect modelling and intelligent support. In our case, it was sufficient for

the environment to provide feedback counts and an analysis of whether the feedback was followed and to expose different feedback types. However, it may be necessary in other cases to modify the actual text provided or provide other adaptations to the whole environment, all based on affect.

Other limitations are due to the methodological choices we had to make. First, to obtain comparative data, human annotated affective states, we used the BROMP methodology to collect observations of student affect during the sessions. This allowed us to achieve ecological validity but reduced the size of the dataset for the evaluation to a subset of the students in the class.

In addition, as with other similar studies in the field, there are reliability and validity issues stemming from the fact that for part of the study we rely on human observations. For example, we were not able to fully hide the experimental condition from the annotators - the setting and whether or not the students were overtly speaking could have made this obvious and implicitly bias the annotations. However, both annotators were trained in the BROMP method, and one of them was not fully aware of the existence of different conditions nor the exact research questions. However, the high agreement between the two annotators gives us confidence in the human observations. There was also promising correlations between the human observations and the system measurements, which again support the validity of the affect detection.

A final limitation, that challenges the internal validity of this study, concerned us both at the design and analysis stages and relates to the direct comparability of our two conditions (*affect* and *non-affect*). For various technical, practical and experimental reasons, these had to differ in many respects, such that ultimately it blurs the aspects that are responsible for the effects that we have reported here. This is a well known issue in evaluating interactive adaptive systems (c.f. Paramythis et al., 2010). Accordingly, we followed a layered process to evaluate different aspects of the learner model, building it on top of previously collected data (Grawemeyer et al., 2015a,b), and we internally validated its robustness. Nevertheless, the *non-affect* condition necessarily restricted other interactive aspects of the system. This was important in our case because we wanted to maintain the ecological validity of the study and to ensure its pedagogical underpinning.

For both ethical and pedagogical reasons, it was better to compare our affective learning approach with the current state of play of interactive learning environments (i.e. those that offer affect-agnostic supportive feedback). But this introduces an experimental confound, because the *affect* condition had additional prompts to the *non-affect* condition. Although we could have let the system provide, for example, prompts based on random affect decisions, we know from our previous studies (Mavrikis et al., 2014; Grawemeyer et al., 2015b) that asking students to speak for the sake of speaking without them perceiving some impact would be problematic. Nevertheless, we consider that this confound does not have an adverse effect on our efforts to determine the potential impact of the affect-aware support since we see the design of the prompts integral to the availability of affect modelling. Notwithstanding that, further research is needed to tease apart any interaction effects.

7 Conclusion

We have developed an adaptive learning environment that provides intelligent formative feedback according to students' affective states. Although affect detection is also used as input for the recommendation of different learning activities (exploratory or structured practice, as described by Janning et al., 2014), in this paper, we focus only on feedback *type* selection and *presentation* as direct manifestations of the system's learner model. Our system includes two Bayesian networks that are able to predict the type and presentation of feedback that has the highest likelihood to improve a student's current affective state. The latter is inferred from a combination of a student's interaction with the learning environment and keywords and prosodic features as they are talking aloud.

We evaluated our *affect*-aware intelligent formative support by comparing it to a *non-affect* version, where feedback was provided based on students' performance. Although our results show only a non-significant difference between the *affect* and *non-affect* conditions on learning gains, the statistically significant increase of knowledge in both conditions and the higher learning gains in the *affect* condition are promising, have implications both for the design of educational technology and for traditional teaching, and warrant further research.

During our evaluation, the students' affective states and task behaviour were annotated by observers. We have shown, this way, that the automatic-detection of the students' affective states correlates highly with the human-detected affective states. Importantly, our results show that in the *affect* condition students were significantly less *bored* than students in the *non-affect* condition. In addition, students in the *affect* condition showed significantly less *off task* behaviour than students in the *non-affect* condition. These are important findings and reaffirm common-sense assumptions, that by responding appropriately to student affect, teachers (not just technologies) are likely to encourage students in a more productive engagement, which in turn will lead to better learning outcomes as boredom and off-task behaviour can have a negative impact in learning.

Future work includes the refinement of the Bayesian networks and the detection of student affective states from keywords, prosodic features and interactions (as shown in Figure 3) with the newly collected data. Also, it would be interesting to compare our automatically-detected and human-detected affective states with student self-reports of their affective states. While the human-detected and self-reports might both have challenges, their triangulation would increase the trustworthiness of the automatic-detection of affective states.We also plan to analyse our data further by looking in depth at the relationship between a student's affective states and interactions with the learning environment, which includes detecting interaction patterns that are associated with particular affective states. In addition, given the promising results of open learner models (OLMs) on learning (c.f. Long and Aleven (2017)), exploring the impact of an open *affective* model would also be especially interesting.

Finally, the platform overall allows for further experimentation that can help tease apart any interaction effects and explore related research questions and hypothesis on the role of affect-aware modelling. This has implications in both the better design of intelligent support systems but also for human teaching in that findings from such a platform can also inform pedagogical strategies for responding appropriately to student affective states.

8 Acknowledgments

This research received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no 318051 - iTalk2Learn project. Thanks to all our iTalk2Learn colleagues for their support and ideas.

References

Acee TW, Kim H, Kim HJ, Kim JI, Chu HNR, Kim M, Cho YJ, Wicker FW (2010) Academic boredom in under- and over-challenging situations. Contemporary Educational Psychology 35(1):17–27

Askeland M (2012) Sound-based strategy training in multiplication. European Journal of Special Needs Education 27(2):201–217

Azevedo R, Witherspoon A, Chauncey A, Burkett C, Fike A (2009) MetaTutor: A MetaCognitive Tool for Enhancing Self-Regulated Learning. In: Proceedings of the AAAI Fall Symposium on Cognitive and Metacognitive Educational Systems, Association for the Advancement of Artificial Intelligence (AAAI) Press., Menlo Park, CA, pp 14–19

Baker RSJd (2007) Modeling and understanding students' off-task behavior in intelligent tutoring systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '07, pp 1059–1068

Baker RSJd, D'Mello SK, Rodrigo MMT, Graesser AC (2010) Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. International Journal of Human-Computer Studies 68(4):223–241

Basu S, Biswas G, Kinnebrew J (2017) Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. User Modeling and User-Adapted Interaction - Special Issue on Impact of Learner Modeling

Bradley MM, Lang PJ (1994) Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry 25(1):49–59

Burleson W, Picard R (2007) Evidence for gender specific approaches to the development of emotionally intelligent learning companions. Special issue on Intelligent Educational Systems, IEEE Intelligent Systems 22(4):62–69

Carenini G, Conati C, Hoque E, Steichen B, Toker D, Enns J (2014) Highlighting interventions and user differences: Informing adaptive information

visualization support. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14, pp 1835–1844

Chi MTH (2000) Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In: Glaser R (ed) Advances in instructional psychology, Mahwah, NJ: Lawrence Erbaum Associates, pp 161–238

Conati C, MacLaren H (2009) Empirically building and evaluating a probabilistic model of user affect. User Modeling and User-Adapted Interaction 19:267–303

Cowie R, Douglas-Cowie E, Apolloni B, Taylor J, Romano A, Fellenz W (1999) What a neural net needs to know about emotion words. Journal of Computational Intelligence and Applications pp 109–114

Csikszentmihalyi M (1990) Flow: The Psychology of Optimal Experience. NY: Harper and Row

D'Mello S, Graesser A (2013) AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. ACM Transactions on Interactive Intelligent Systems 2(4):1–38

D'Mello S, Lehman B, Sullins J, Daigle R, Combs R, Vogt K, Perkins L, Graesser A (2010) A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning. In: Intelligent Tutoring Systems: 10th International Conference, ITS 2010

D'Mello SK, Graesser AC (2010) Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. User Modeling and User-Adapted Interaction 20(2):147–187

D'Mello SK, Kory J (2015) A review and meta-analysis of multimodal affect detection systems. ACM Computing Surveys 47(3):43:1–43:36

D'Mello SK, Craig SD, Gholson B, Franklin S, Picard RW, Graesser AC (2005) Integrating affect sensors in an intelligent tutoring system. In: Affective Interactions: The Computer in the Affective Loop Workshop at the International Conference on Intelligent User Interfaces, pp 7–13

D'Mello SK, Lehman B, Pekrun R, Graesser AC (2014) Confusion can be beneficial for learning. Learning & Instruction 29(1):153–170

Epp C, Lippold M, Mandryk RL (2011) Identifying emotional states using keystroke dynamics. In: Annual Conference on Human Factors in Computing Systems, pp 715–724

Forbes-Riley K, Litman D (2011a) Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. Speech Communication 53(9-10):1115–1136

Forbes-Riley K, Litman D (2011b) Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. Computer Speech & Language 25(1):105–126

Grawemeyer B, Mavrikis M, Hansen A, Mazziotti C, Gutiérrez-Santos S (2014) Employing speech to contribute to modelling and adapting to students' affective states. In: Proceedings of the 9th European Conference on Technology Enhanced Learning, EC-TEL 2014, Springer International Publishing, Lecture Notes in Computer Science, pp 568–569

Grawemeyer B, Holmes W, Gutiérrez-Santos S, Hansen A, Loibl K, Mavrikis M (2015a) Light-bulb moment? towards adaptive presentation of feedback based on students' affective state. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15, ACM, New York, NY, USA, pp 400–404

Grawemeyer B, Mavrikis M, Holmes W, Hansen A, Loibl K, Gutiérrez-Santos S (2015b) Affect matters: Exploring the impact of feedback during mathematical tasks in an exploratory environment. In: Proceedings of the 17th International Conference on Artificial Intelligence in Education, AIED 2015, Springer International Publishing, Lecture Notes in Computer Science, pp 595–599

Gutiérrez-Santos S, Mavrikis M, Magoulas G (2012) A separation of concerns for engineering intelligent support for exploratory learning environments. Journal of Research and Practice in Information Technology 44(3):347–360

Hattie J, Timperley H (2007) The Power of Feedback. Review of Educational Research 77(1):81–112

Hayes AF (2013) Introduction to mediation, moderation, and conditional process analysis: A regression-based approach. Guilford Press.

Holmes W, Mavrikis M, Hansen A, Grawemeyer B (2015) Purpose and Level of Feedback in an Exploratory Learning Environment for Fractions. In: Conati C, Heffernan N, Mitrovic A, Verdejo MF (eds) Artificial Intelligence in Education, Springer International Publishing, Lecture Notes in Computer Science, pp 620–623, URL http://link.springer.com/chapter/10.1007/978-3-319-19773-9_76, dOI: 10.1007/978-3-319-19773-9_76

Janning R, Schatten C, Schmidt-Thieme L (2014) Feature analysis for affect recognition supporting task sequencing in adaptive intelligent tutoring systems. In: Proceedings of the 9th European Conference on Technology Enhanced Learning, EC-TEL 2014, Springer International Publishing, Lecture Notes in Computer Science, pp 179–192

Janning R, Schatten C, Schmidt-Thieme L (2016) Perceived task-difficulty recognition from log-file information for the use in adaptive intelligent tutoring systems. International Journal of Artificial Intelligence in Education 26(3):855–876

Jaques N, Conati C, Harley JM, Azevedo R (2014) Predicting affect from gaze data during interaction with an intelligent tutoring system. In: Proceedings of the 12th International Conference of Intelligent Tutoring Systems, ITS 2014, Springer International Publishing, Lecture Notes in Computer Science, pp 29–38

Jiang D, Cui Y, Zhang FP X, Ganzalez I, Sahli H (2011) Audio visual emotion recognition based on triple-stream dynamic bayesian network models. In: Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction, Springer International Publishing, Lecture Notes in Computer Science, pp 609–618

Kaliouby RE, Robinson P (2004) Real-time inference of complex mental states from facial expressions and head gestures. In: Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops

2004

Kirschner P, Sweller J, Clark RE (2006) Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational Psychologist 41(2):75–86

Kort B, Reilly R, Picard RW (2001) An affective model of the interplay between emotions and learning. In: Proceedings of the IEEE International Conference on Advanced Learning Technologies, ICALT '01, IEEE Computer Society, 43-46

Lang PJ, Greenwald MK, Bradley MM, Hamm AO (1993) Looking at pictures: Affective, facial, visceral, and behavioral reactions. Psychophysiology 30(3):261–273

Long Y, Aleven V (2017) Enhancing learning outcomes through self-regulated learning support with an open learner model. User Modeling and User-Adapted Interaction - Special Issue on Impact of Learner Modeling

Mavrikis M, Maciocia A, Lee J (2007) Towards predictive modelling of student affect from web-based interactions. In: Proceedings of the 2007 Conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work, IOS Press, Amsterdam, The Netherlands, The Netherlands, pp 169–176, URL http://dl.acm.org/citation.cfm?id=1563601.1563632

Mavrikis M, Geraniou E, Noss R, Hoyles C (2008) Revisiting pedagogic strategies for supporting students' learning in mathematical microworlds. In: Proceedings of the International Workshop on Intelligent Support for Exploratory Environments at EC-TEL, vol 8, pp 41–50

Mavrikis M, Gutiérrez-Santos S, Geraniou E, Noss R (2013) Design requirements, student perception indicators and validation metrics for intelligent exploratory learning environments. Personal and Ubiquitous Computing 17(8):1605–1620

Mavrikis M, Grawemeyer B, Hansen A, Gutiérrez-Santos S (2014) Exploring the potential of speech recognition to support problem solving and reflection - wizards go to school in the elementary maths classroom. In: Proceedings of the 9th European Conference on Technology Enhanced Learning, EC-TEL 2014, Springer International Publishing, Lecture Notes in Computer Science, pp 263–276

Mazziotti C, Holmes W, Wiedmann M, Loibl K, Rummel N, Mavrikis M, Hansen A, Grawemeyer B (2015) Robust student knowledge: Adapting to individual student needs as they explore the concepts and practice the procedures of fractions. In: Workshop on Intelligent Support in Exploratory and Open-ended Learning Environments Learning Analytics for Project Based and Experiential Learning Scenarios at the 17th International Conferenece on Artificial Intelligence in Education, AIED 2015, pp 32–40

Nasoz F, Alvarez K, Lisetti CL, Finkelstein N (2003) Emotion recognition from physiological signals for presence technologies. International Journal of Cognition, Technology and Work, Special Issue on Presence 6(1):4–14

Ocumpaugh J, Baker RSJd, Rodrigo MMT (2012) Baker-Rodrigo Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. Tech. rep., New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

Paleari M, Benmokhtar R, Huet B (2009) Evidence theory-based multimodal emotion recognition. In: Proceedings of the 15th International Multimedia Modeling Conference, MMM 2009, Springer Berlin Heidelberg, Lecture Notes in Computer Science, pp 435–446

Paramythis A, Weibelzahl S, Masthoff J (2010) Layered evaluation of interactive adaptive systems: framework and formative methods. User Modeling and User-Adapted Interaction 20:383–453

Pekrun R (2006) The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review 18(4):315–341

Piaget J (1951) Organization and pathology of thought: Selected sources. In: Principal factors determining intellectual evolution from childhood to adult life, Columbia University Press, New York, NY, US, pp 154–175

Porayska-Pomsta K, Mavrikis M, Pain H (2008) Diagnosing and acting on student affect: the tutor's perspective. User Modeling and User-Adapted Interaction 18(1):125–173

Rowe JP, Mott BW, McQuiggan SW, Robison JL, Lee S, Lester JC (2009) CRYSTAL ISLAND: A narrative-centered learning environment for eighth grade microbiology. In: Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education, AIED 2009, pp 11–19

Rummel N, Mavrikis M, Wiedmann M, Loibl K, Mazziotti C, Holmes W, Hansen A (2016) Combining Exploratory Learning with Structured Practice to Foster Conceptual and Procedural Fractions Knowledge. In: Proceedings of the 12th International Conference of the Learning Sciences, ICLS 2016, pp 58–65

Sail-Labs (2016) SAIL LABS Technology GmbH. URL `http://www.sail-labs.com`

Santos OC, Saneiro M, Salmeron-Majadas S, Boticario JG (2014) A methodological approach to elicit affective educational recommendations. In: Proceeding of the 14th International Conference on Advanced Learning Technologies, ICALT 2014, pp 529–533

Schuller B, Müller R, Lang M, Rigoll G (2005) Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensemble. In: Proceedings of the 9th European Conference on Speech Communication and Technology, Interspeech 2005, pp 805–808

Shen L, Wang M, Shen R (2009) Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment. Educational Technology & Society 12(2):176–189

Shute VJ (2008) Focus on Formative Feedback. Review of Educational Research 78(1):153–189

del Soldato T, du Boulay B (1995) Implementation of motivational tactics in tutoring systems. Journal of Artificial Intelligence in Education 6(4):337–378

Sweller J, van Merrienboer JG, Paas GW (1998) Cognitive Architecture and Instructional Design. Educational Psychology Review 10:251–296

Vail AK, Boyer KE, Wiebe EN, Lester JC (2015) The Mars and Venus Effect: The Influence of User Gender on the Effectiveness of Adaptive Task Support. In: Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization, UMAP 2015, pp 265–276

Vanlehn K (2006) The behavior of tutoring systems. International Journal of Artificial Intelligence in Education 16(3):227–265, URL `http://dl.acm.org/citation.cfm?id=1435351.1435353`

VanLehn K, Burleson W, Girard S, Chavez-Echeagaray E, Gonzalez-Sanchez J, Hidalgo-Pontet Y, Zhang L (2014) The affective meta-tutoring project: Lessons learned. In: Proceedings of the 12th International Conference on Intelligent Tutoring Systems, ITS 2014, pp 84–93

Vogt T, André E (2005) Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: Proceedings of the 2005 IEEE International Conference on Multimedia and Expo, ICME 2005, pp 474–477

Vyzas E, Picard RW (1998) Affective pattern classification. In: Proceedings of the AAAI Fall Symposium Series: Emotional and Intelligent: The Tangled Knot of Cognition, pp 23–25

Wöllmer M, Metallinou A, Eyben F, Schuller B, Narayanan SS (2010) Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH '10, pp 2362–2365

Woolf B, Burleson W, Arroyo I, Dragon T, Cooper D, Picard R (2009) Affect-aware tutors: recognising and responding to student affect. International Journal of Learning Technology 4(3-4):129–164

9 Vitae of authors

– **Dr Beate Grawemeyer**
Birkbeck, University of London, Department of Computer Science and Information Systems, BBK Knowledge Lab, Malet Street, London WC1E 7HX, UK
Dr Beate Grawemeyer is a Postdoctoral Researcher at Birkbeck, University of London, UK. She has an MSc in Knowledge-Based Systems and a PhD in Computer Science and Artificial Intelligence, both from the University of Sussex. She has worked in several areas of artificial intelligence and human-computer interaction. Her primary research interest lies in the development of user / learner models that offer novel adaptation techniques (such as

adaptations to affective states) to enhance user experience and performance when conducting specific learning tasks. The paper reflects this interest.

– **Dr Manolis Mavrikis**
UCL Institute of Education, University College London, UCL Knowledge Lab, 23-29 Emerald Street, London WC1N 3QS, UK
Dr Manolis Mavrikis is a Reader (Associate Professor) in Learning Technologies at UCL Knowledge Lab. He holds a BSc in Mathematics from University of Athens, Greece with an emphasis in education, MSc with distinction in Informatics and PhD in Artificial Intelligence in Education from the University of Edinburgh. His research interests developed over more than 15 years of experience, lie at the intersection of learning sciences, human-computer interaction and artificial intelligence. Manolis?s research centres on employing learning analytics to help teachers, schools, education ministries or researchers develop an awareness and understanding of the processes involved in learning, and on designing evidence-based intelligent technologies that provide direct feedback to learners, such as the work presented in this paper.

– **Dr Wayne Holmes**
The Open University, Institute of Educational Technology, Milton Keynes MK7 6AA, UK.
Dr Wayne Holmes is a Lecturer (Assistant Professor) in the Institute of Educational Technology at the Open University, UK. He received his PhD (DPhil) in Education (Learning and Technology) from the University of Oxford and has degrees in Film (BA), Philosophy (MA) and Education (MSc Oxon). Previously, as a researcher at the UCL Knowledge Lab, University College London, he contributed to the iTalk2Learn project where he was responsible for the development of the intervention model and a novel approach to formative feedback. Recently, he co-authored the report "Intelligence Unleashed - An Argument for Artificial Intelligence in Education".

– **Dr Sergio Gutiérrez-Santos**
Birkbeck, University of London, Department of Computer Science and Information Systems, BBK Knowledge Lab, Malet Street, London WC1E 7HX, UK
Dr Gutierrez-Santos received his BEng and MEng in Electronic Engineering from University Carlos III of Madrid, where he also got his PhD in 2007. He also pursued studies in Philosophy at National Distance Education University (UNED, Spain) and Philosophy of Mind and Logic at Birkbeck. Since 2012 he has been a Lecturer at the Department of Computer Science at Birkbeck. His research focuses on user modelling and knowledge representation, especially in the context of exploratory learning systems such as those used in the research described in this paper.

– **Dr. Michael Wiedmann**
Robert Bosch Stiftung GmbH, Heidehofstr. 31, 70184 Stuttgart, Germany
Dr. Michael Wiedmann is Senior Project Manager for education and digitalization at Robert Bosch Stiftung, a major German foundation that

has managed the philanthropic bequest of entrepreneur Robert Bosch for over 50 years. Dr. Wiedmann received his diploma and PhD in psychology from University of Freiburg. In his role as a researcher in the Educational Psychology Lab in the Institute of Educational Research at Ruhr-Universität Bochum, Germany, he contributed to the summative evaluation in the iTalk2Learn project. He is interested in technology-enhanced learning, teacher training, and competence assessment.

– **Prof. Dr. Nikol Rummel**
Ruhr-Universitt Bochum, Institute of Educational Research, Universitätsstrasse 150, 44801 Bochum, Germany
Dr. Nikol Rummel is a Full Professor and head of the Educational Psychology Lab in the Institute of Educational Research at Ruhr-Universität Bochum, Germany. She is also an Adjunct Professor in the Human-Computer Interaction Institute at Carnegie Mellon University, Pittsburgh, USA. Dr. Rummel is the current (2016-17) president of the International Society of the Learning Sciences (ISLS). She is Associate Editor of the Journal of the Learning Sciences, and Editorial Board member of the International Journal of Computer-Supported Collaborative Learning, of the International Journal of Artificial Intelligence in Education, and of Learning & Instruction. Her research published in numerous journal articles and book chapters focuses on developing and evaluating instructional support for learning in computer-supported settings, with an emphasis on CSCL and on adaptive learning support, in particular. The paper presented in this volume thus closely ties in with her main research interest.